DOCUMENT RESUME

ED 244 952 SP 024 450

AUTHOR Duckett, Willard R., Ed.
TITLE Observation and the Evaluation of Teaching. A CEDR Monograph.
INSTITUTION Phi Delta Kappa, Bloomington, Ind. Center on Evaluation and Research.
PUB DATE [80]
NOTE 133p.
AVAILABLE FROM Phi Delta Kappa, P. O. Box 789, Bloomington, IN 47402.
PUB TYPE Collected Works - Conference Proceedings (021) -- Reports - Evaluative/Feasibility (142) -- Information Analyses (070)

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS *Classroom Observation Techniques; Elementary Secondary Education; Evaluation Criteria; *Evaluation Methods; Higher Education; Inservice Teacher Education; Measures (Individuals); *Research Utilization; Student Teacher Relationship; Teacher Behavior; Teacher Characteristics; Teacher Effectiveness; *Teacher Evaluation; *Values

ABSTRACT

The wide variety of contexts in which good teaching takes place, a desirable circumstance from most points of view, creates problems, not the least of which is evaluation. Systematic observation is one measurement procedure that can provide some of the data essential to reliable observation. The fourteenth National Symposium for Professionals in Evaluation and Research (NSPER) focused on the development of observation scales, checklists, and rating scales as well as other topics. This report of the three NSPER:80 sessions is a record of the presentations; in addition, each article has been revised to incorporate ideas that emerged from two and one half days of discussion at each site. Contents of this publication include: (1) "Classroom Observation: Potential and Problems" (Thomas L. Good); (2) "Research on Teaching as a Basis for the Evaluation of Teaching" (John A. Zahorik); (3) "Qualitative Versus Quantitative Approaches to Observing Teaching" (John A. Zahorik); (4) "Values as a Basis for the Evaluation of Teaching" (John A. Zahorik); (5) "What Research Says about the Components of Teaching" (Larry A. Braskamp); (6) "Components of Teaching Based on Values and Valuing" (Larry A. Braskamp); (7) "Guidelines for Using Rating Scales and Checklists" (Larry A. Braskamp); (8) "Techniques for Systematic Observation" (Robert B. Ingle); (9) "Developing Rating Scales and Checklists" (Robert B. Ingle); and (10) "Components of a Humane System for Evaluating Teaching" (William J. Gephart). (JMK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Observation and the Evaluation of Teaching

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D. Kliewer

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

A CEDR Monograph
PHI DELTA KAPPA

NSPER '80

Observation and the Evaluation of Teaching

Edited by

WILLARD R. DUCKETT
Phi Delta Kappa

Phi Delta Kappa
P.O. Box 789
Bloomington, IN 47402

A Publication of Phi Delta Kappa's
Center on Evaluation, Development and Research

Cover Drawing by Paula Ahmad

Table of Contents

PREFACE ..... v.

NSPER:80 — TALLAHASSEE
Classroom Observation: Potential and Problems
Thomas L. Good 2

NSPER: 80 — PHOENIX
Research on Teaching as a Basis for the Evaluation of Teaching
John A. Zahorik 46

Qualitative Versus Quantitative Approaches to Observing Teaching
John A. Zahorik 51

Values as a Basis for the Evaluation of Teaching
John A. Zahorik 56

NSPER: 80 TOLEDO
What Research Says About the Components of Teaching
Larry A. Braskamp 62

Components of Teaching Based on Values and Valuing
Larry A. Braskamp 87

Guidelines for Using Rating Scales and Checklists
Larry A. Braskamp 97

NSPER: 80 — CO-DIRECTORS
Techniques for Systematic Observation
Robert B. Ingle 106

Developing Rating Scales and Checklists
Robert B. Ingle 112

Components of a Humane System for Evaluating Teaching
William J. Gephart 117

Preface

The mandate is clear. We want quality teaching in our schools. But what is *quality* teaching? Is it the activities of the firm disciplinarian who brooks no nonsense as he or she marches students through a variety of activities enroute to mastery of some subject matter? There are many who would say yes. Others would say that good teaching creates an environment that continually challenges students and stimulates their growth. Neither our values nor our research has yet identified specific characteristics of teachers, or classroom behaviors, that adequately describe good teaching. We are fairly clear, however, on one point: Teaching is relative to contexts, and good teaching occurs in a variety of contexts.

The wide variety of contexts in which good teaching takes place, a desirable circumstance from most points of view, creates problems, not the least of which is evaluation. If the contexts for high quality teaching vary so much, how can we measure or evaluate them all? How can we get the data or evidence we need to identify the conditions that consistently result in good teaching?

Systematic observation is one measurement procedure that can provide some of the data essential to reliable evaluation. Systematic observation — including observation checklists and rating scales — is essentially a process of selective watching. It is guided by a list of things to look for as the teacher goes about his or her classroom activities. With carefully developed observation instruments such as checklists or rating scales, we can gather information that will be useful to teachers who want to improve their teaching skills and will also be useful to those who have the responsibility for making personnel decisions about teachers.

NSPER:80 (the fourteenth National Symposium for Professionals in Evaluation and Research) focused on the development of observation scales, checklists, and rating scales. The topic also covered: 1) the components of good teaching that have been established by research, 2) other components of good teaching that evolve from the values of the people who will use the observation instrument, 3) identification of typical activities and behaviors to be observed, 4) ranking the importance of the characteristics or behaviors to be observed (again, based on the values of the people who use the instruments), and 5) guidelines for using systematic observation in evaluating teachers.

The major presenter at each of the three NSPER:80 sessions has had extensive experience in the observation of teachers, and

the co-directors have organized and conducted dozens of workshops and symposiums on the evaluation process. Participants in the three sessions included practitioners with a wide variety of experience and expertise related to the evaluation of teachers. The NSPER:80 schedule was structured to promote discussion between the resource people and the participants, discussions about common problems and solutions.

This report of the three NSPER:80 sessions is more than a record of the presentations; each article has been revised to incorporate ideas that emerged from two and a half days of discussion at each site. The authors have attempted to inform the reader about the various contexts in which evaluation of teaching can occur and to guide one in the development of observation instruments that are "custom made" for the settings in which they will be used.

Systematic observation of teaching is not easy, but it can be crucial to the improvement of evaluation and, thus, very beneficial for the improvement of teaching itself.

Willard Duckett
Phi Delta Kappa

NSPER: 80

Tallahassee Presentation

THOMAS L. GOOD
University of Missouri
at Columbia

Classroom Observations: Potential and Problems

Thomas L. Good

In 1970 there was little systematic observational data available to describe classroom events. In 1969 Schwab noted that we had little data to describe normative teaching practice in American schools. It was impossible to answer relatively simple questions about life in classrooms (e.g., Do teachers vary on how much time they allocate for mathematics instruction?). Despite the fact that no observational data existed, both school critics and reformers frequently advocated simple universal prescriptions for improving American education. At times, critics of American education have called for more openness and at other times for more structure. Indeed, some critics have argued both positions at different times.

In 1980, after a decade in which much classroom observational data was collected in reasonably sophisticated and rigorous fashion, we now have much more information about some aspects of classroom behavior. Observational data offer no panacea to education problems, because teaching is a problematic activity. However, recent observational data yield important information about schooling and provide a beginning base upon which subsequent studies can build.

In this paper I want to point out some of the gains that have occurred from 1970s research which utilized systematic observational procedures and suggest some of the practical advantages that can occur if school districts use observational procedures in appropriate ways. The potential of observational procedures to help practitioners conceptualize their classroom behavior, and to encourage teachers to consider alternative ideas about teaching, is very high. However, if not used judiciously, observational data can result in overcontrol and overprescription of classroom teaching, and hence do more harm than good. In this paper I will discuss common problems associated with classroom observation and make some suggestions for improving observations. However, it is beyond the scope of this paper to provide detailed comment on how to *conduct* classroom observations appropriately. This topic has been developed elsewhere in the book *Looking in Classrooms* (Good and Brophy, 1978). The interested reader can find detailed information there.

Important Information from Observational Research in the 1970s

Observational research in the 1970s provides solid evidence that *teachers make an important difference in student learning.* Although on the basis of their own experience, both as students and observers in classrooms, many practitioners have long contended that teachers make an important difference, many critics of schools argue that teachers make little, if any, difference in the educational attainments of students. Indeed, much research concerning schools in general has provided data that have been interpreted by some individuals as suggesting that teachers do not affect student learning.

Findings by Coleman et al. (1966) have often been interpreted as suggesting that schooling does not make an important difference in educational attainment. However, the problem with this research and many other studies is that it did not provide observational measures of what individual teachers did in the classroom. Systematic data collected in the 1970s suggest that teachers vary widely in their classroom behavior, and some research using correlational descriptive procedures followed by field experiments provides solid evidence that certain patterns of teacher behavior make important differences in student learning (e.g., Anderson, Evertson, and Brophy, 1979; Good and Grouws, 1979). Because of such observational research, there is now solid evidence that teachers teaching similar pupils under similar circumstances often behave in different ways, and some of these differences in teacher behavior can be associated with differential learning gains. Furthermore, recent research has illustrated that some of the techniques used by effective teachers can be taught to other teachers who then use the new methods successfully (e.g., Good and Grouws, 1979). It is beyond the scope of this paper to describe the methodology utilized in these studies or to report the substantive findings extensively. However, the interested reader can consult Good (1979), Brophy (1979), Peterson and Walberg (1979), and Rosenshine (1979) for more information.

School critics often describe schools in pejorative terms and claim that teachers behave in very *similar* ways. Simple descriptions encourage many reformers to provide relatively simple prescriptions for improving classroom teaching and learning. There may be a number of dimensions upon which American schools and teachers can be criticized; however, uniformity of teacher behavior is one factor upon which school critics are flatly wrong.

That is, teachers exhibit an extraordinarily vast number of differences in teaching behavior and technique. This fact has been illustrated repeatedly by observational research in the 1970s.

Rosenshine (1980) and Fisher et al. (1978) provide data from the Beginning Teacher Evaluation Study to show that teachers vary greatly in the amount of time they spend on particular subject areas in elementary schools. Indeed, because a student is assigned to a particular teacher, he or she may spend twice as much time (or more!) studying mathematics than a student who is assigned to a different teacher in the same school. In addition to the amount of time that teachers spend on a particular subject, there is also a great variance in how teachers *use* that time. For example, Good and Grouws (1977) found that some teachers spend a great deal of time teaching mathematics by assigning seatwork; however, other teachers spend somewhat more time on the development portion of the lesson (explaining the meanings of concepts and how to do work) prior to making seatwork assignments.

Classrooms vary not only in how teachers distribute time and present subject matter to students, but also in how they organize classrooms for learning. Some classrooms are taught basically in whole groups, others in two or three working groups, and other classrooms provide a great deal of individualization. In some classrooms students are encouraged to work together and to cooperate with others; in other classrooms teachers are students' primary source of feedback and evaluation. Formal studies of classrooms (e.g., Bossert, 1979) and reviews and conceptualizations of classroom life (Doyle, 1979) have revealed that classrooms which often appear similar do in fact place quite different demands upon students, and that other classrooms appear dissimilar but place corresponding learning demands upon students. Hence, there is now ample data to suggest that teachers both act and structure their classrooms in highly different ways. Although many potential teaching-learning interactions do not appear in classrooms, certainly many variations in classroom practice currently exist.

To reiterate, variation in teacher behavior has been related to student achievement. Good and Grouws (1979) have found that certain patterns of instructional behavior in mathematics relate to student achievement, and Anderson, Evertson, and Brophy (1979) identified instructional acts in reading group instruction which are associated with student achievement. Stallings (1980) has presented experimental data to show that *how teachers utilize time* can affect the reading progress of students in secondary classrooms.

Classroom observation also reveals that program labels do not predict teacher or student behaviors or student progress. For example, some classrooms which are labeled "open" provide more structure and direct teaching than classrooms designated as traditional. Study after study has revealed that teachers vary widely in how they implement a particular "program." Similarly, many studies and reviews have illustrated that the "method" is not as important as the *quality of teaching* (Chall, 1967). That is, some teachers who use individualized methods obtain reasonably good student achievement, whereas other teachers use whole-class methods and obtain equally good results. There are some teachers who have comparatively little positive impact upon student achievement who use both methods. The organizational *form* of instruction makes little difference independent of the quality with which the form is implemented. It does appear that some forms of instruction, such as large group instruction featuring active teaching, appear to be easier to implement effectively. When implemented effectively, such strategies are associated with student gains in basic skill areas. Still, it is not difficult to find classrooms where large group instruction is done poorly and where students achieve at comparatively low levels.

Classroom observational research in the 1970s demonstrated that classrooms are varied and quite complex. Such findings serve to negate the popular misconception that there are *universal forms of instruction* that tend to work for all students across all subjects and in virtually any school context. Instructional problems vary from classroom to classroom, and there are no universal observational systems or formulas that can be used for describing and/or improving classroom instruction. For example, some classrooms are "undermanaged," while others are "overmanaged." Encouraging all teachers to structure classrooms more fully would lead some teachers who already have an appropriate degree of structure to overstructure their classrooms. Although classroom observational research is beginning to yield some usable information and concepts, this information must be applied in a decision making format that takes into consideration a particular teacher's strengths and goals as well as the general interests and abilities of the students in that class.

Observational research has also revealed that teachers often behave in ways that are *inconsistent with what is believed to be emphasized in teacher education programs.* For example, Brophy (1981) provides data indicating that teachers praise considerably less than would be expected, considering the emphasis placed upon the utility of praise in many classroom methodology textbooks.

Table 1*

Guidelines for Effective Praise

EFFECTIVE PRAISE	INEFFECTIVE PRAISE
1. Is delivered contingently	1. Is delivered randomly or unsystematically
2. Specifies the particulars of the accomplishment	2. Is restricted to global positive reactions
3. Shows spontaneity, variety, and other signs of credibility; suggests clear attention to the student's accomplishment	3. Shows a bland uniformity, which suggests a conditional response made with minimal attention
4. Rewards attainment of specified performance criteria (which can include effort criteria, however)	4. Rewards mere participation, without consideration of performance processes or outcomes
5. Provides information to students about their competence or the value of their accomplishments.	5. Provides no information at all or gives students information about their status
6. Orients students towards better appreciation of their own task-related behavior and thinking about problem solving	6. Orients students toward comparing themselves with others and thinking about competing

7. Uses students' own prior accomplishments as the context for describing present accomplishments
8. Is given in recognition of noteworthy effort or success at difficult (for *this* student) tasks

9. Attributes success to effort and ability, implying that similar successes can be expected in the future
10. Fosters endogenous attributions (students believe that they expend effort on the task because they enjoy the task and/or want to develop task-relevant skills)
11. Focuses students' attention on their own task-relevant behavior

12. Fosters appreciation of and desirable attributions about task-relevant behavior after the process is completed

7. Uses the accomplishments of peers as the context for describing students' present accomplishments
8. Is given without regard to the effort expended or the meaning of the accomplishments (for *this* student)
9. Attributes success to ability alone or to external factors such as luck or easy task

10. Fosters exogenous attributions (students believe that they expend effort on the task for external reasons to please the teacher, win a competition or reward, etc.)
11. Focuses students' attention on the teacher as an external authority figure who is manipulating them
12. Intrudes into the ongoing process, distracting attention from task-relevant behavior

---

*This table is reprinted from the work: Brophy, J. "Teacher Praise: A Functional Analysis." *Review of Educational Research*, vol. 51, 1981, pp. 5-32.

Brophy contends that infrequent praise is probably best in most settings, because he found that teachers often use praise in ways that are highly inappropriate. Table 1 illustrates appropriate and inappropriate aspects of praise and suggests the careful preparation and thinking that has to occur before a concept like praise can be adequately measured in the classroom.

Research on teacher planning has also revealed that popular notions about how teachers should plan are not the strategies that teachers use when faced with pressures and demands of actual classroom teaching.

One of the most surprising findings of observational work resulted from the research of Kounin (1970) and others who have subsequently used his concepts and ideas. At one time educators believed that how misbehavior was dealt with was very important to teaching. Subsequent studies found that regardless of how teachers reacted to misbehavior, instructional time was lost for many students and that the key to good management was *preventing* problems. Research in the 1970s revealed several strategies that successful managers use to minimize needless interruptions or delays. Many studies on classroom management have been conducted at the Research and Development Center for Teacher Education by Carolyn Evertson and colleagues.

It would be possible to illustrate other advantages occurring from observational research in the 1970s; however, the examples above should sufficiently make the point that observational research has begun to yield useful concepts and findings. Although it is beyond the scope of this paper to provide methodological details, it is important to emphasize that these gains in knowledge, through use of observational procedures, have resulted not only because of more observation during research but also because better instruments were utilized, better sampling techniques were employed in selecting teachers for study (often selecting only teachers who had demonstrated an ability to improve student learning over a period of time), and because in some instances observational procedures were more sensitive to the perceptions of teachers and students.

Non-Research Uses of Observational Procedures

Despite the important gains that have resulted from the successful use of sophisticated, systematic observational procedures and other research activities, there has been comparatively little

sign of widespread improvement in the use of observational procedures in normal school improvement/inservice activities. This is too bad, because appropriate use of observational procedures offers much potential for understanding and for possibly improving classroom events. Indeed, I know of no strategy that has as much potential for improving classrooms.

INSERVICE TEACHER OBSERVATION

There are many ways in which classroom observation might be utilized successfully by school systems in addition to the "normal" use of observation for evaluation purposes. For example, school districts might use classroom observational procedures to raise expectations of particular teachers and/or beginning teachers. This might be particularly important in urban school districts in which new teachers have not had any experience interacting with students who come from home situations that are radically different from their own. After a brief period in which teachers hold and reject unrealistically high expectations about their efficacy in the classroom, many teachers may begin to withdraw psychologically from such a teaching situation and begin to subconsciously blame students when learning does not take place. The opportunity to visit teachers who are having positive impact and success with disadvantaged students would be a reasonable way to help teachers see that there are successful ways of involving such students in learning activities.

Observation would also allow teachers to observe different classrooms and to visit with teachers who use teaching styles that differ from their own. Opportunities to observe other teachers would help eliminate the physical isolation and separation that are often associated with teaching and could encourage teachers to develop more tolerance for different instructional strategies. Too often teachers utilize only one set of teaching methods (because the teacher that they did their student teaching with taught in a particular way) and are hesitant to try other methods. Observation may help teachers to become aware that almost any method can work or fail. The key to successful teaching is found in 1) the quality of thinking and teaching that accompanies a particular classroom method, and 2) the impact of the teaching on student thinking and performance. Success is not dependent on whether lecture, small-group work, or some other singular method is used.

Another possible gain from classroom observation is coordination of instructional programs within a school. Although there is no reason to expect, or to require, uniform practices of teaching in a school, occasionally the discrepancy between how one teacher presents a particular subject and how teachers in the following grades present the subject is so great that it is likely to cause some students difficulty. For example, if problem-solving in mathematics in third grade is taught basically as production of a single, best solution (taking a long time to form a working hypothesis), and in the fourth grade it is taught as trying to produce as many solutions as possible, students may become confused about what they are trying to do in mathematics. They may begin to view the subject as an arbitrary set of rules and become relatively passive in their attempts to learn. I am not suggesting that teachers ought to behave in standard ways, but vast discrepancies in an instructional program ought to either be modified or explained to students (last year we emphasized a particular approach...this year we are emphasizing a different strategy for the following reasons...).

Finally, and most importantly, observational data may encourage teachers to think more seriously about how their actions influence students and about how students reciprocally influence teacher behavior. Being observed may encourage teachers to take their role as teacher more seriously and attempt a variety of ways to improve student learning, *particularly if they have some control over what is observed.*

TEACHER AWARENESS

Perhaps the most important reason that observation can be useful is that teachers are not aware of many aspects of their classroom behavior. Parents have often lamented that when they ask their son or daughter, "What did you do today?", the child shrugs his or her shoulders and says, "I don't know." If a similar question was asked of teachers (e.g., "What did you say to Judy or Tom today?"), one might very well receive a similar response, at least for many students in the classroom. This is because classroom behavior occurs very rapidly and many events occur simultaneously. Jackson (1968) has estimated that elementary school teachers engage in over 1,000 interactions with students each day. In secondary classrooms, teachers may have as many as 150 students to deal with during an instructional day. Hence, to be able to remember the frequency of inter-

actions with particular students and the quality of each of those interactions is virtually an impossible task.

In addition to the number and frequency of interactions that take place in the classroom, a teacher also has to make rapid decisions. For example, if the teacher hears a noise in the back of the room but does not see the student who made the noise, should the teacher call down a student who is in the vicinity and who frequently engages in misbehavior? When a student asks the teacher a question for information about a particular assignment, is the student asking a serious question and seeking help or is the student simply attempting to waste time or make the teacher look foolish? When the teacher asks a student a question and the student responds inappropriately, is it the teacher's fault? (Was the question asked inappropriately or awkwardly?) Was it the student's fault? (Has the student not paid attention or not read the assignment?) The way in which the teacher answers these questions of *causality* determines subsequent behaviors of the teacher. In making such decisions, the teacher has to include not only evaluations related to the student (e.g., the general ability level of the student, the apparent involvement of the student at present), but also the general context in which the event is occurring (e.g., Is the teacher running out of time and several general comments have to be made before the class ends? Is the class becoming restless?) Hence, it is easy to see how teachers may behave in quite different ways towards students and not be aware of those differences because decisions are made so rapidly and in complex situations.

Teachers might easily form different expectation levels for students because so much behavior in the classroom is ambiguous. Is a student's question a real one or is the student merely seeking attention? Does the raised eyebrow on the face of the student mean that he or she is thinking, or needs help? Teachers may inadvertently develop higher expectations for certain students and systematically interpret *ambiguous* behavior in positive ways. At the same time, they may interpret equally ambiguous behavior in negative ways when coming from students for whom they have low expectations. Regardless of the causes of differential behavior, it is clear that teachers do behave in distinct ways toward individual students. Jackson and Lahaderne (1967) have collected data which indicate that some students receive so little teacher contact in the classroom that for all practical purposes they might as well be in a classroom of 100 as in a classroom of 30. Other students in classes of only 30 students receive so much interaction with the teacher that psy-

chologically they enjoy a private seminar with the teacher. Elsewhere, Jere Brophy and I (Good and Brophy, 1980) have described a number of ways in which teachers vary in their contacts with students whom they believe to be of high or low achievement potential by:

1. seating low expectation students far from the teacher or seating them in a group,
2. paying less attention to lows in academic situations (e.g., smiling less often and maintaining less eye contact),
3. calling on lows less often to answer classroom questions or to make public demonstrations,
4. waiting less time for lows to answer questions,
5. not staying with lows in failure situations (e.g., providing fewer clues and asking fewer follow-up questions),
6. criticizing lows more frequently than highs for incorrect public responses,
7. praising lows less frequently than highs after successful public responses,
8. praising lows more frequently than highs for marginal or inadequate public responses,
9. providing lows with less accurate and less detailed feedback than highs,
10. failing to provide lows with feedback about their responses as often as highs,
11. demanding less work and effort from lows than highs, and
12. interrupting performance of lows more frequently than highs.

In follow-up work, after observational data were collected showing that teachers varied in their behavior toward students perceived to be high and low in potential, it was determined that teachers were *unaware* of many of these differences (Good and Brophy, 1974). Specifically, teachers were much more likely to be unaware of the qualitative aspects of their behavior (how they interacted with students) than how frequently they interacted with particular students (although some teachers were largely unaware of both qualitative and quantitative aspects of behavior).

Other investigators have reached similar conclusions. Goodlad and Klein (1974) found that there is often a large discrepancy between teachers' perceptions of their own innovative behavior and the perceptions of observers. Many teachers truly thought they were individualizing instruction and encouraging inductive learning when, in fact, observers found the instruction to be quite structured and dominated by teachers. Others have also commented on the discrepancy not only between teachers'

opinions and those of observers, but also on differences between teacher and student perceptions (e.g., Cooper and Good, in press; Hook and Rosenshine, 1979).

The example used in this section has stressed teacher awareness of behavior toward students believed to be of high or low achievement potential. The list of differences in how some teachers treat students perceived as high or low achievers would be a reasonable way to organize classroom observation. However, there is a virtually unlimited number of topics that could be explored in the classroom. Other examples include teachers' 1) awareness of their behavior toward boys and girls, 2) awareness of the behavior that boys and girls exhibit toward them, and 3) awareness of the types of expectations teachers communicate to students when introducing and ending lessons. These and many other dimensions of classroom behavior that teachers may be unaware of are discussed in Good and Brophy (1978).

WHAT TO OBSERVE

There is no single set of variables or a given observational instrument that will be useful or important in all classroom settings. Clearly, an observational system should be selected or designed in order to provide relevant information about teaching rather than manipulating teaching to fit the observational instrument. There are three basic situations in which an observational instrument might be useful: 1) implementation of a program developed outside a school, 2) exploration of ideas suggested by research, and 3) testing both teachers' and others' ideas about improving practice.

If a school district wants to adopt a particular instructional program, it may want to collect evidence to determine whether the program is in fact being implemented and whether students' responses (e.g., achievement) are influenced when successful implementation occurs. This need appears to be especially important, considering recent research evidence suggesting that teachers who are believed to be implementing a "program" often teach in vastly different ways.

Program implementation. In my own research I have worked with school districts which are interested in implementing the Missouri Mathematics Program. I used correlational research (observing how relatively "effective" and "ineffective" teachers presented mathematics in the classroom) and subsequent

treatment research (encouraging other teachers to behave as relatively more effective teachers had been observed to behave). Doug Grouws and I have identified a set of teaching behaviors that appear to be associated with general achievement gains in the classroom (Good et al., 1978). In particular, we have found that teachers who emphasize the conceptual meaning of the lesson through a careful development lesson, and who carefully check to see if students have comprehended and understood the concepts and skills being presented in the development lesson, have students who work much more successfully during independent seatwork. These students move through the curriculum more quickly than do students in classrooms where teachers do not emphasize the development portion of the lesson and do not engage in systematic follow-through (reteaching as necessary) and systematic review. Hence, if a school district is attempting to implement the Missouri Mathematics Program, it needs to be able to identify the extent to which it is being followed in the classroom. In Tables 2-4 some of the instruments used for this assessment are presented.

In working with the Missouri Mathematics Program, we have found that students in experimental groups have consistently achieved at higher levels than students in control groups. However, the purpose here is not to advocate this program of teaching as the best way to present mathematics in upper elementary and junior high school years, but rather to illustrate one purpose of observation. Unfortunately, programs are too often adopted without any observational check to see if the program is being taught in the way in which it was designed.

Schools' ideas. Other reasons for observations may involve the implementation of a school program that has been designed by the school staff. For example, teachers may be concerned about the lack of prosocial attitudes that students hold toward one another and begin to encourage students to work with one another during parts of the instructional day. If teachers are to assess the short-term effects of their strategy, it would be very useful to have an observational instrument that can describe how students spend their time when interacting with one another. It is conceivable that certain patterns of interaction are more successful than others. For example, a teacher who allows students to work together in small groups might want to assess whether students were learning to respect and to listen to one another, or perhaps it would indicate that small group work was intensifying negative feelings that students held toward one another. (If that teacher were using small group work to facil-

Table 2

Treatment Program Description

Junior High Study

The treatment involves the following:

1. The beginning portion of the lesson (first 12 minutes) should have three parts:
    a. A brief review,
    b. Checking of homework, and
    c. 3-5 minutes of mental computation exercises.
2. Approximately 10 minutes daily should be spent on verbal problem solving.
3. Approximately 20 minutes daily should be devoted to developing conceptual understanding of mathematics.
4. Approximately 30 minutes worth of homework should be assigned each night except Friday.
5. Approximately 10-15 minutes of seatwork should be given daily to provide uninterrupted successful practice.
6. The first 20 minutes each Monday should be spent on review and the full period should be used for review every fourth Monday.
7. The teacher should get everyone involved in the seatwork immediately and keep them involved.
8. The teacher should make the students accountable for the seatwork.

Check List

Junior High Mathematics Study

_____ Monday ? Date _____ Teacher_____

Observer _____ Observation # _____

Place a check mark (✓) in the blank next to all statements that hold true.

INTRODUCTORY PHASE

1. _____ A brief review of previous work was conducted in a meaningful context.
2. _____ 2-5 minutes was spent checking homework.
3. _____ 1-5 minutes was spent on mental computation.
4. _____ There was a slow transition to the main part of the lesson.

DEVELOPMENT

5. _____ Development moved too slow.
6. _____ Development moved too fast.
7. _____ Teacher paced development with progress of students.
8. _____ The students were held accountable for controlled practice during the development phase.
9. _____ The teacher spent 15-21 minutes developing the mathematics in the lesson (disregarding verbal problem solving).

SEATWORK

10. _____ Teacher spent 10-20 minutes on seatwork.
11. _____ The teacher ensured that students were actively engaged in seatwork during the first 1½ minutes.
12. _____ The teacher was available to provide immediate help and actively supervised student seatwork.
13. _____ Students were accountable for the seatwork at the end of the seatwork phase.
14. _____ Seatwork directions took longer than one minute.
15. _____ The teacher assigned homework.

VERBAL PROBLEM SOLVING

16. _____ The teacher spent approximately 10 minutes on verbal problem solving.
17. _____ The teacher spent the majority of the period on verbal problem solving.
18. _____ The following verbal problem-solving techniques were used.
    - _____ Problems without numbers.
    - _____ Students were asked to create and write verbal problems.
    - _____ Students estimated the answers to verbal problems.
    - _____ Attention was given to the reading of verbal problems.
    - _____ Students were asked to write open sentences.

Table 3

Description of Lesson Components to be Coded
Junior High Mathematics Study

The following is a description of each of the phases of the class period which will be coded. Whenever the teacher's behavior fits the particular description, code it as follows:

DEVELOPMENT

The development portion of the class period is that part of the lesson devoted to establishing comprehension of skills and concepts. Development relates almost exclusively to work with *new* ideas, concepts and skills. Activities used during this phase often include teacher explanations and demonstrations, and may include the use of manipulative materials, concrete examples, making comparisons and searching for patterns, class discussions, group work, and the use of audio-visual materials.

Often during some part of development, the teacher will pose oral questions to students to assess their comprehension of the topic at hand. The teacher may also use controlled practice during this phase. That is, one or two problems are given at a time and then immediate feedback is given on the correctness of responses. This is usually done while the teacher is trying to identify and correct student misunderstandings. A brief summary of the prerequisite ideas and skills necessary to do or understand the topic of the day is considered part of development. Note that its focus is to facilitate the development portion of the lesson and not simply to refresh past skills and concepts.

(continued)

PRACTICE SEATWORK (PS)

Seatwork refers to practice work which students complete individually at their desks. Seatwork refers to written work only and does not include oral practice.

PRACTICE BOARDWORK (PB)

Boardwork refers to practice work that students complete individually at the board. Like seatwork, this work relates primarily to work with *new* ideas, concepts, skills and objectives which were presented in the development phase of that day's class period.

REVIEW (RE)

Review refers to work on *old* material. It deals with concepts which the students have had prior to the particular day in question (whether it be material from the previous day, or from much earlier). Use of games, puzzles, worksheets, etc., which are used to review skills and ideas, fall into this category.

MENTAL COMPUTATION (MC)

Mental computation is computation that is done without the aid of pencil and paper (or minicalculator). The processing is done mentally. The problems which the students are asked to solve mentally may relate either to new or to old ideas.

(continued)

HOMEWORK (HW)

This phase includes all activities involving homework. Thus, it includes the teacher reading off the correct answers to the homework, the in-class grading of homework, the showing of solutions to homework problems, and/or the collecting of homework papers.

TRANSITION (TR)

Transition refers to the process of going from one phase of the class period to another. Transitions should be coded only if they are noticeable (i.e., involve a minute or more, or are not done smoothly). Transition also involves the period of time from when the class is scheduled to begin and when it actually gets productively under way.

VERBAL PROBLEM SOLVING (VP)

Verbal problem solving refers to the time the teacher spends working on word problems. This phase often includes the teacher demonstrating problem solving strategies, solving problems, having the students estimate answers to verbal problems, and/or having the students solve verbal problems.

NAME: SCHOOL: WEEKS BEGINNING &

Put a check (✓) in the columns that apply and fill in other appropriate information.

	STRATEGY USED													
	BRIEF REVIEW	HOMEWORK CHECKED	MENTAL COMPUTATION	ABOUT 20 MINUTES ON DEVELOPMENT	ABOUT 10 MINUTES ON VERBAL PROBLEMS	PROBLEMS WITH-OUT NUMBERS	WRITING VERBAL PROBLEMS	ESTIMATING THE ANSWER	READING VERBAL PROBLEMS	WRITING OPEN SENTENCES	SEATWORK PAGE/PROB. #	HOMEWORK PAGE/PROB. #	COMMENTS*
Monday													
Tuesday													
Wednesday													
Thursday													
Friday													
Monday													
Tuesday													
Wednesday													
Thursday													
Friday													

*Please note review days, test days, assemblies, shortened periods, quizzes, etc.

Table 4*

Description of High Inference Scales

Junior High Mathematics Study

CLARITY

Clarity refers to the degree to which the teacher's presentation of material and his substantive interactions with students are understood by them.

5 *Very high clarity*. The teacher's explanations are easy to understand and pupil questions are adequately answered. The teacher seems aware of the pupil's levels, sensing problems they are having or may have.

4 *High clarity*. Between moderate and very high.

3 *Moderate clarity*. The teacher seems to be understood by most pupils, but not all of the time. Sometimes the teacher is confusing and vague.

2 *Low clarity*. Between very low and moderate.

1 *Very low clarity*. Pupils seem very confused by the presentation. The teacher cannot answer the pupils' questions, or answers them in an unclear manner by using concepts and terms the pupils are apparently unfamiliar with or by being overly complex and ambiguous.

ENTHUSIASM

This scale is used to judge the extent to which the teacher displays interest, vitality, and involvement in his subject and his instruction.

5 *Very high enthusiasm*. The teacher is stimulating, energetic, and very alert. He seems interested and involved in what he is teaching; moves around, gestures, inflects voice.

(continued)

4 *High enthusiasm*. Between moderate and very high.

3 *Moderate enthusiasm*. Occasionally the teacher seems interested and involved; some display of activity, such as gesturing. Sometimes the teacher is dull, routine, and lacking in vigor.

2 *Low enthusiasm*. Between very low and moderate.

1 *Very low enthusiasm*. The teacher's behavior is lethargic, dull, routine; a minimum of vocal inflection, gesturing, movement, or change in facial features. The teacher appears to lack interest in what he is doing.

•The clarity and enthusiasm scales are based upon the work by Emmer, E., "Classroom Observation Scales" (mimeo). Research and Development Center for Teacher Education, University of Texas: Austin, Texas 1967.

MANAGERIAL

This scale is used to judge the degree of effectiveness of the managerial skills displayed by the teacher.

5 *Very high management*. The teacher is an effective manager. She structures, maintains and monitors learning activities. She runs the class with a minimum of disruptions.

4 *High management*. Between moderate and very high.

3 *Moderate management*. Occasionally the teacher is an effective manager; some display of structuring, maintaining and monitoring learning is present. Sometimes she manages the room very ineffectively and allows too many disruptions.

2 *Low management*. Between very low and moderate.

1 *Very low management*. The teacher manages the classroom very inefficiently and ineffectively. Too many disruptions are tolerated.

ACCOMPLISHMENT INDEX

5 *Very high accomplishment*. The teacher accomplishes a remarkable amount during the class period in terms of the number of examples used, the number of problems worked, the amount of material covered, and so on, in relation to what seemed possible.

4 *High accomplishment*. Between moderate and very high.

3 *Moderate accomplishment*. At times the teacher seems to be accomplishing a lot but at other times things drag with very little being accomplished, in relation to what seemed possible.

2 *Low accomplishment*. Between very low and moderate.

1 *Very low accomplishment*. The teacher accomplishes very little compared to what seemed possible.

INTERACTION INDEX (excluding seatwork)

5 *Very high participation*. The teacher's behavior patterns elicit a large amount of *active* student participation and interaction. The students willingly and enthusiastically take an *active* part in the lesson. There are a lot of self-initiated questions and teacher questions. Frequent question and answer sessions are observed.

4 *High participation*. Between moderate and very high.

3 *Moderate participation*. Occasionally there is some *active* student participation and interaction, but at other times there is very little.

2 *Low participation*. Occasionally there is some *active* student participation and interaction, but at other times there is very little.

1 *Very low participation*. There is little or no *active* student participation and interaction during the lesson.

(continued)

IMPLEMENTATION

Use this scale to indicate how well the teacher implemented the prescribed treatment behaviors (circle one). See the attached page for a description of the major parts of the treatment program.

5	4	3	2	1
Implemented all major components of the program	Implemented most of the major components	Implemented about one-half of the program	Implemented some of the program components	Implemented very little of the program

OVERALL QUALITY

Describe the teacher's overall quality based on the entire class period.

5	4	3	2	1
Excellent	Very Good	Average	Fair	Poor

itate achievement and not prosocial gains, then the teacher might look for different outcomes.)

Observational research has shown again and again that changes in instructional *form* do not automatically lead to improvements in teaching. Attempts to improve classroom situations are often not successful and may even make a situation worse. By use of systematic observation, one can determine whether the pattern of behaviors hoped for is actually being realized and whether these patterns are having positive or negative effects upon desirable outcomes that are anticipated when changes are made in an instructional program.

Adaptation of research. Individual schools or school systems can also use concepts that are developed in observational research to examine classroom behavior. That is, the schools are not adapting a program but rather using concepts developed by research to examine or change instructional behaviors.

An example of a district level program using observation as a key aspect is the work of Martin (1973). Mary Martin and Sam Kerman in the Los Angeles County school district were intrigued with some of the results that Jere Brophy and I obtained about how some teachers differentiated their behavior toward high- and low-achieving students (some of these behavioral differences were presented earlier). Building upon ideas that we had presented in the first edition of *Looking in Classrooms* (1973), these authors designed a training program that helped to sensitize teachers to possible ways in which they might unconsciously interact with low-achieving students in ways that made their continuing low achievement quite likely.

An interesting aspect of their program was that teachers had the chance to observe each other—that is, participating teachers observed others who were teaching similar students under like circumstances. The teachers were well versed in the observation system as it directly corresponded with teaching techniques emphasized in the training program. That is, the observational instrument allowed teachers to describe ways in which other teachers interacted with a few high- and a few low-achieving students in the classroom. Students of teachers who had the opportunity to participate in the program exhibited higher achievement and attitude scores than did students of teachers in control classrooms. Hence, the training and observational format were successful in improving student attitudes and achievement.

Teachers reported that the opportunity to observe in other teachers' classrooms was helpful, and they considered observation

an important professional activity. Teachers reported that they learned many new alternative ways of thinking about teaching, and that they were motivated by observing other teachers interacting successfully with students described as low achievers. That is, once teachers saw other teachers having successful interactions with low-achieving students, they were motivated to find ways consistent with their own teaching style in which they could have similar effects on students.

Research thus provides yet another basis for classroom observation. Ideas that have emerged from research can be used as a basis for describing and thinking about classroom behaviors. The example above stressed the application of concepts and observational techniques coming from observational research on teachers' expectations. Much of the recent research on teaching effectiveness (Brophy, 1979; Good, 1980), classroom management (Anderson, Evertson, and Emmer, in press; Brophy and Putnam, 1979), and time utilization (Denham and Lieberman, 1980) has yielded concepts and programs around which meaningful inservice programs could be developed.

Observation as an evaluative tool. Whether the ideas result from research or whether they are completely developed from the teacher's own understanding and experience, the ideas need to be tested and demonstrated through actual observation. For example, the teacher who is concerned that many of his or her students do not approach and ask for clarification when they are not sure about how to proceed on an independent seatwork assignment may attempt to do two or three things to improve the situation. The teacher may provide clearer and more detailed procedural directions and may end the classroom discussion or recitation period with an opportunity for students to get information. "I may have given the information too quickly to you, so if there are any problems in understanding what you are to do, please ask now. It's okay; I'm waiting for any questions that you may have about the assignment."

Although such teaching behavior is designed to help students comprehend the assigned task, such an emphasis on clarity and providing psychological conditions for encouraging students to ask questions may *not* be successful. Students may remember other teachers who have said, "feel free to ask questions," but who consistently provided pejorative feedback to students foolish enough to ask questions. Alternatively, the change in teacher behavior may result in too much change in students' behavior. They may begin to ask needlessly and frequently for clarification of minor issues that they could figure out themselves.

Observational instruments need to be modified. It is therefore important to know what one is trying to effect by changing his or her classroom behavior and to look for signs of improvement. For example, a teacher might want to check for indications of more work involvement during the seatwork period and a higher percentage of students handing in completed assignments at the end of the period. For many observational procedures to be successful, they have to be relevant to the particular goals of the teacher who is being observed. Furthermore, observational instruments often have to be constructed or revised so that they can be responsive to a particular teacher's needs. For example, the teacher might legitimately ask a principal or supervisor for help with the following problem: "My students often complain that my examinations are unfair and are too difficult and that although they study for them, their performance on the exam does not reflect how well they thought they understood the material. What can I do?" Although there are many observational systems that have been developed, no system I know of would be helpful for answering this question unless it was modified.

Key considerations in answering the question of what to observe are the goals that a particular teacher or school district have developed and teaching problems that develop. To reiterate, whether the ideas for observation are coming from research and theory or from the practical insights and wisdom of practitioners, it is important that the observational instrument be appropriate for the context in which the data are being collected. For example, even though a school district may be concerned about equity issues (e.g., do high- and low-achieving students receive fair amounts of teacher time and attention?), the way in which that question is addressed by an observational form may vary somewhat from classroom to classroom. If a particular teacher spends 80% of his or her time holding discussion/recitation sessions, it would make sense to raise questions about the extent to which the teacher calls upon students or allocates response opportunities and provides feedback to individual students during public interactions. However, if the teacher spends 80% of the day allowing students to work on individual tasks, it might be more important to determine when and under what circumstances students approach the teacher for feedback. It would also be important to know how decisions are made about the *content* that particular students work on (Is the work too easy and boring for certain students? When can they move from one assignment to another?). Similarly, many beginning teachers

need observation focused upon their general classroom organization and management skills, whereas experienced and reasonably successful teachers would benefit from more specialized observation and feedback. Just as administrators ask teachers to adapt instruction to students when necessary, administrators need to adapt observational procedures to fit the needs of individual teachers.

Problems with Observation

In this section I want to describe a few problems that are associated with the use of observational procedures in schools for staff development purposes. One of the basic difficulties is that observational instruments and procedures are adopted for use much too casually. A state develops an instrument for evaluating teachers and suddenly other states become interested in this particular instrument. A researcher produces a new instrument and others become interested in using that system without understanding its purpose or limitations. If observation is to be effective, supervisors, principals, and teachers all need to be actively involved in determining *why* observation will take place and developing the type of coding instruments that are needed at a particular time (what is needed one year may not be needed the following year).

A second problem with observation as utilized in public schools is that most evaluation procedures and observational instruments are too simple (e.g., they measure only a very few of the many possible behaviors that are occurring in the classroom). In addition, these instruments are rule-based. That is, some observational procedures assume that certain patterns of behavior are effective, independent of how they influence student learning. For example, at one time the conceptualization that teachers ought to be more indirect was popular, and many observational systems have been developed to examine the extent to which teachers and students talk in the classroom. There are abundant training materials for teaching teachers how they can obtain more student talk. Such simple-minded, rule-based thinking is hopelessly narrow and self-defeating. Unless observation is used to describe and to think about teaching, it is not worth the time and effort to collect data. Surely there are times when teacher talk should prevail and times when students should be more responsive. Simple codes that only assess the frequency of responses mask important information; they do not reveal the quality or appropriateness of the talk, for example.

The context in which a given student or teacher behavior occurs has to be taken into account when the "appropriateness" of a behavior is being judged. For example, teachers can err by correcting behavior at the wrong time or in an inappropriate fashion. Table 5 presents a form, taken from the book, *Looking in Classrooms*, that illustrates a way to code the *context* of behavior as well as its frequency. The reader may also want to reexamine Table 1 as it also indicates that *how* a teacher behaves (praises) is at least as important as how frequently a teacher performs a particular behavior.

Yet another problem inherent in the way in which many schools use observational procedures is the tendency to *overinterpret* a particular behavior and to make assessments too swiftly. Most classroom teachers can relate personal experiences of receiving feedback from individuals who have misinterpreted what was occurring in their classrooms. Such individuals have offered useless advice because there was too little intent to understand or to think about the particular behaviors that were occurring.

I know of one situation in which an art supervisor had categorically told all first-grade teachers in the school district that under *all* circumstances, when children were doing art work, they should be using a big magic marker and not a pencil. On one occasion the principal in a particular school building was observing a first-grade classroom and saw a student using a pencil during an art activity conducted by the regular classroom teacher. The principal immediately and publicly asked the teacher why the student was using a pencil when the art supervisor had asked that pencils not be used. Before the teacher could respond, the student indicated that he was making *spaghetti* and that he was well aware of the fact that he was not to use a pencil during art activities, but for the life of him, he could not understand how he could make spaghetti with a thick line emitted by the magic marker! Subsequently, the art supervisor assured the principal that he had done the right thing and that only magic markers should be used!

Unfortunately, stories about the misuse of observational data are all too common in the experience of classroom teachers. Part of the reason that observers tend to overinterpret and react too quickly is that they are poorly trained and probably have inappropriately high expectations for the simplistic and immediate results that can be obtained through observational procedures. Observation is hard work, and it takes trained individuals if it is to yield intelligent descriptions about classroom behavior that

Table 5

Teacher's Reactions to Inattention and Misbehavior[1]

USE: When the teacher is faced with problems of inattention or misbehavior

PURPOSE: To see if teacher handles these situations appropriately

Code the following information concerning teacher's response to misbehavior or to inattentiveness. Code only when teacher seems to be aware of the problem; do not code minor problems that teacher doesn't even notice.

BEHAVIOR CATEGORIES		CODES		
		A	B	C
A. TYPE OF SITUATION				
1. Total class, lesson, or discussion	1.			
2. Small group activity-problem in group	2.			
3. Small group activity-problem out of group	3.			
4. Seatwork checking or study period	4.			
5. Other (specify)	5.			
B. TYPE OF MISBEHAVIOR	6.			
1. Brief, nondisruptive, should be ignored	7.			
2. Minor, but extended or repeated. Should be stopped nondisruptively	8.			
3. Disruptive, should be stopped quickly. No questions needed	9.			
4. Disruptive, questions needed or advisable	10.			
5. Other (specify)	11.			
	12.			

C. TEACHER'S RESPONSE(S)

1. Ignores (deliberately)
2. Nonverbal; uses eye contact, gestures or touch, or moves near offender
3. Praises someone else's good behavior
4. Calls offender's name; calls for attention or work; gives rule reminder. No overdwelling
5. Overdwells on misbehavior, nags
6. Asks rhetorical or meaningless questions
7. Asks appropriate questions-investigates publicly
8. Investigates privately, now or later
9. Threatens punishment if behavior is repeated
10. Punishes (note type)
11. Other (specify)

13. _____ _____ _____
14. _____ _____ _____
15. _____ _____ _____
16. _____ _____ _____
17. _____ _____ _____
18. _____ _____ _____
19. _____ _____ _____
20. _____ _____ _____
21. _____ _____ _____
22. _____ _____ _____
23. _____ _____ _____
24. _____ _____ _____
25. _____ _____ _____

CHECK IF APPLICABLE

_____ 1. Teacher delays too long before acting, so problems escalate
_____ 2. Teacher identifies wrong student or fails to include all involved
_____ 3. Teacher fails to specify appropriate behavior (when this is not clear)
_____ 4. Teacher fails to specify rationale behind demands (when this is not clear)
_____ 5. Teacher attributes misbehavior to ill will, evil motives
_____ 6. Teacher describes misbehavior as a typical or unchangeable trait; labels student

---

[1]This form is reprinted from the book *Looking in Classrooms*, 2nd Ed., by Good, T. and J. Brophy. New York: Harper & Row, 1978.

may eventually promote improved patterns of behavior. Many principals take only limited course work in curriculum and instruction (if any) when they work on advanced degrees and have had little training in the use of observational instruments. Without such background information and experience, it is impossible for a principal to provide *instructional leadership* unless he or she is willing to develop expertise in collecting and communicating the results of observational research.

Earlier in the paper, I argued that teachers may be unaware of many things that happen in the classroom because events occur so rapidly and because teachers have to interpret ambiguous events quickly (Good and Brophy, 1978; Rohrkemper, in press). The same difficulties also confront a classroom observer who must understand what to look at, when, and for how long. Many behaviors that an observer must react to will be ambiguous and are apt to be overinterpreted unless the observer is well trained. Hence, just as teachers may selectively spend more time with certain students and not be aware of it, observers may tend to overrate or underrate teachers because of their general perceptions of them, not because of the specific instructional behaviors they observe. That is, a poorly trained observer may rate a teacher high in all aspects of a behavior because the observer is favorably impressed with the teacher in terms of performance on one or two aspects of classroom behavior.

Another reason classroom observation is often misused is that not enough observations are collected to develop an adequate sample of how teachers and students behave over time. The fact that teachers are rarely observed in many schools helps to explain why principal or supervisor ratings of teachers appear to relate poorly with other descriptions of teacher effectiveness. All teachers have good and bad days. Some teachers have proportionately more good days than bad; however, all teachers will have at least some average or mediorcre days. Decisions and estimates based on instructional behavior observed on only one or two occasions are untenable assumptions. If one cannot make several observations, it is likely that no observation should be made, because no reliable and usable information can be obtained from only one or two observations.

Another problem is overreliance on one particular system. Earlier, the point was made that school personnel are often too cavalier in selecting an observational instrument for use in a particular classroom or in a set of classrooms. However, even if school personnel spend a great deal of time selecting a system that seems useful at a particular time, the continued use of that

system over a long period of time will produce diminishing returns. As a case in point, the Georgia Performance Assessment System appears to be a reasonably interesting system upon which one might want to make three or four observations during one particular year. However, to continue to use that system would appear to be self-defeating because the instrument is largely a survey instrument that enables one to look at several aspects of teaching behavior. However, as teachers are observed to be reasonably good in some areas and need development and work in other areas, other observational systems that measure more *detailed aspects* of particular types of teacher behavior in specific contexts would be necessary to bring about improvements in teaching and learning. It may therefore be practical to use an instrument for a year or so, but to continue to use that instrument year after year seems to be inherently self-defeating in most cases.

Another common problem with school-based observational procedures is that many important *proactive* teacher behaviors are ignored. Teacher thinking and behaviors that are important for producing more and less desirable forms of student thinking and behavior may have occurred the day before (or several days before) an observation; hence, observers may overreact to what they see on a particular day. For example, observers may note that the review session of the lesson goes very poorly. When a teacher asks students questions about concepts and ideas that were presented on the previous day and students respond poorly, observers may look for reasons in the present situation that explain the inadequate performance. They may be tempted to suggest different styles or ways of asking questions that might produce more desirable student responses. However, it may be that the inadequate performance of the students is due not to how the teacher conducts the review session, but to how the teacher handled the developmental portion of the lesson the day before.

Although observational systems can often help teachers become more aware of unprofitable or unexamined classroom behavior, there are many instances in which classroom observation will prove to be an inadequate strategy. Observation must therefore be seen as a valuable but limited method for understanding classroom behavior (as are all methods when they are used alone). There are some behaviors that simply cannot be explained in terms of the present, ongoing situation by an outsider who has missed most of the interactions that have occurred in a classroom earlier in the year. As a case in point, I know of an

incident in which a curriculum supervisor advised a teacher that a shy student who was sitting in the rear corner of the classroom should not be there and that the teacher had probably isolated the student because she didn't want to deal with the student. In actuality, the teacher changed seating patterns three times a month, and this was one of the first times during the year that the student had expressed a strong preference and the teacher had honored it. Unfortunately, the need to provide advice and to be helpful is too strong in some supervisors who overreact to particular events or who feel that they must be of immediate help to teachers. To reiterate, observation raises questions and suggests possibilities; it does not necessarily provide immediate answers.

In addition to the problem of not being able to witness classroom events that have transpired earlier during a school year, experience of teachers and students in preceding years may control classroom behavior. For example, certain students may be uncomfortable and unwilling to approach a teacher not necessarily because of how that teacher structures or reacts to feedback situations, but because of experiences these students have had with other teachers in the past. Observers need to be aware that not all events can be explained by observational procedures and that often observational data need to be supplemented with teacher and student interviews.

It is also important to realize that students' perceptions of classroom events vary on some occasions from those of their teachers and, indeed, some students may react to the same teacher behavior in quite different ways. For example, in observational work conducted with Jere Confrey, I recently observed a teacher who told students to write a paragraph with a "stubby" pencil and a "crinkly" sheet of paper and then to *think* about the paragraph and to rewrite it with a clean sheet of paper and an ink pen. However, *the teacher did not explain* to the students what he meant by the word "think," and *students did not ask* the instructor what he meant by the term and what he expected them to do when they rewrote the paragraph. In interviews with the instructor, it became clear that he wanted students to pay attention to content/substantive issues when rewriting their paragraphs (e.g., choice of verb). However, in interviews with students, most, but not all, students felt that the instructor was interested in "form" (legible handwriting, neatness) rather than content. Classroom events are difficult for teachers to interpret, but they are equally difficult and ambiguous for observers and students to understand. It is

comparatively easy for any classroom participant to project his or her own sense of reality onto what is taking place, unless there is a concerted effort to reliably collect information and describe events rather than try to evaluate or interpret them immediately.

It must be acknowledged and recognized by those who use observational systems that *any* observational procedure has inherent value judgments associated with it. Some observational instruments are much more favorable toward structured teaching, whereas other instruments record more events that are associated with informal teaching. Hence, teachers whose styles differ from the instruments being utilized to observe them are likely to be unduly penalized, even though they may be teaching effectively within the styles that they have chosen to use. Some instruments emphasize affective goals, whereas other instruments place more stress upon cognitive goals. Practitioners need to realize that observation is not value-free and, although some observational procedures and coding systems are more descriptive than others, the types of variables that are examined and the general orientation that an observational system provides are likely to be better suited for certain values, goals, and teaching styles than for others.

Similarly, observers bring biases to the classroom and those biases can become self-fulfilling when observers attempt to evaluate or to describe teaching behavior. The observer who likes well-managed and highly structured classrooms may inadvertently react more positively to teachers who exhibit structure, even in classrooms that appear to be over-structured (i.e., classrooms that appear to be too rigid and unbending in the eyes of observers who have different perspectives). Similarly, observers who have high tolerance for noise, flexibility, and ambiguity may give unnecessarily high marks to a teacher who has constructed an open learning environment that many students in the classroom find too noisy, distracting, and fragmented. As has been pointed out elsewhere (Good and Brophy, 1978), observers who become aware of their biases and who understand them have taken the first step toward being able to objectively think about and react to teachers who present styles that are different from their own preferences.

Improving Observation

As has been noted previously in this paper, there are many observational systems that a school district can use to increase opportunities for teachers to receive feedback and to benefit

from it. Perhaps the first step that a school district needs to take is to examine a wide variety of sources and instruments that can be used and to consider seriously the different types of codes that are available, both in terms of format for collecting the information and the substantive issues that the coding system addresses. Teachers should be consulted about issues that are of concern to them. Teachers, supervisors, and others involved in inservice training should read books that address relevant issues and provide a variety of illustrations about what and how to code in classrooms. Obviously, as a co-author of *Looking in Classrooms* (Good and Brophy, 1978) I am biased, but I think this book provides a wealth of ideas about the mechanics of coding, as well as information about how to establish a climate and rapport so that useful and developmental feedback can be shared among teachers. Others have written about classroom observation and many of these sources are excellent as well (e.g., Stallings, 1977). School districts that are serious about utilizing observational procedures effectively need to read several such books as they begin inservice planning.

In addition to involving teachers in selecting and learning about observational procedures, teachers should be involved in the actual creation of new instruments and the selection of available instruments that will be utilized for a particular period of time in certain classrooms. On occasion different instruments will be needed in different classrooms, and instruments will have to be varied over time. A unique instrument does not necessarily need to be used in every classroom. Indeed, it may be that 75% of a coding sheet will be common across teachers and related to important school goals. However, it is important that the form of the observation be applicable to different styles of teaching and that there be some information collected that is of unique interest to a particular classroom teacher.

Unfortunately, there is often an orientation in school practices that can be characterized in the following sequence of behavior: Define it, look at it, then forget about it. Unfortunately, because of the dynamic and problematic nature of teaching, we cannot expect to come up with a single ideal observational system, or to examine teacher behavior on one or two occasions and make some permanent decision that does not have to be examined again. If schools are to use observational procedures in effective ways, there has to be a long-term commitment to the process, and teachers have to become involved in sharing and using information (discussed in Good and Brophy, 1978).

Another way to improve observational procedures is for teachers and curriculum supervisors/observers to take brief refresher courses in child development and learning theory that are appropriate to particular teaching contexts. As we gather more classroom data, it becomes increasingly clear that the context in which teaching occurs is an important part of any statement that attempts to characterize effective teaching (Brophy and Evertson, 1976; Brophy, Evertson, Anderson, Baum, and Crawford, 1981). Simply put, certain behaviors that "work" in first-grade reading groups may not be important in seventh-grade mathematics teaching. Unfortunately, teachers in preservice university settings often do not know the particular context in which they will teach. They may know that they will teach in an elementary school, but whether that will be a suburban first-grade classroom or an inner-city sixth-grade classroom depends upon many factors over which the teachers have little or no control. However, once a teacher is established in a particular context, then it becomes possible to use information about how students learn and develop in much more dynamic ways. The learning needs of second-grade students are different in many ways from those of seventh-grade students because of differences in cognitive and physical maturation. Hence, inservice work using observational procedures can relate more meaningfully to the context in which teachers are teaching if such information is used appropriately. Recent activities by Amarel (1981) suggest that teachers can become more active observers as they become more sensitive to the differences in the ways in which students attempt to learn and integrate information.

If observational procedures are to be improved, several observers should be utilized in three or four consecutive grade levels rather than having individual observers study an entire school district. That is, it would appear to make much more sense for certain observers to work in K-3 grade classrooms, other observers in 4-7 grade classrooms, and yet other observers in 8-12 grade classrooms. In addition to becoming specialists in observational procedures and techniques and ways for sharing that information with teachers, observers could become more attuned to the developmental and motivational needs of students at a particular age. They could also become more familiar with the content that is being presented in particular grade levels. Observation does not yield automatic results. Observers have to be well trained in the mechanics of observation, but they also need knowledge about the content at a grade level and characteristics of children in a particular context.

To yield information that lends itself more readily to decision making, it is often appropriate to observe in a particular class for three or four consecutive days. Too often, school-based consultants model researchers who randomly sample teacher behavior and who visit a teacher several times over the entire year (i.e., it may be several days, weeks, or even a month between visits). Although there are times and reasons to visit but periodically, consecutive visits will often yield more insights than will occasional, random visits.

Perhaps the most important step in improving observation is for all concerned to realize that observational information does not yield quick solutions to educational problems. It may help to describe the problems and sources of interference more clearly and to suggest *possible* solutions, but it does not provide answers. Such information yields tentative suggestions and hypotheses about how to proceed. Unfortunately, after describing a problem many observers present solutions to teachers in relatively compelling terms and, again unfortunately, many teachers take this advice completely or they reject it completely. What is needed is the use of observational data in a decision-making format so that observers and teachers share information about what is taking place in the classroom, collectively begin to think about alternative procedures that might be helpful, and make plans for collecting more evidence about the extent to which alternative plans improve classroom instruction.

Part of the reason for my personal enthusiasm for "patterning" observational visits (observing on three or four consecutive days during the mathematics and/or reading portion of the lesson) is due to some of our own recent experience in trying to characterize effective teaching. Essentially, I have found that classroom observers are often accurate in predicting teachers who are getting relatively poor achievement gains. However, they appear to have a great deal of difficulty identifying those teachers who are getting average gains and those who are getting very high gains from their students. The inference that I make from such data is that identifying something that is not working well is comparatively easier than determining when something is working so-so versus very well. It seems that effective teaching is not so much a masterfully compelling performance as it is a plausible and consistent set of behaviors that integrate learning experience across time. Observers who can observe a lesson being developed over several days can then evaluate the extent to which the teacher communicated the purpose for the learning

activity, and how the activity actually corresponded with the purposes that the teacher had stated. Observing for several days in a row will allow one to see the extent to which teachers build upon previous lessons and use concepts that students have already learned and whether teachers include appropriate review and opportunities for successful practice. Also, observation over several days can offer the observer the opportunity to look for discrepancies in communication that may be occurring between teachers and students. It seems to me that the opportunity to visit a class for several days in a row would allow one to make a more wholistic assessment of what is taking place in terms of the total instructional program—from beginning instruction through the final evaluation and feedback activities.

In many school districts teachers are visited only once or twice a year, and many veteran teachers are not visited at all! Under these circumstances, it is ridiculous to even talk about possible advantages of observation. In terms of normative practice, the recommendation that one spend several consecutive days with a teacher may represent a very high expectation, but an expectation that is nevertheless appropriate. More and more school districts are becoming interested in using observational procedures to improve instruction. For school districts that are serious about these possibilities and are working to develop effective plans for implementing observational procedures, I recommend that if individual teachers can only be visited six or seven times a year, that the school district visit the teacher on two occasions for three or four consecutive days rather than spread those six visits out over the entire year. Obviously, if one has resources to use for observational purposes, the particular strategy for utilizing those resources will vary, depending upon the circumstances in a particular school. For example, if a high school principal is spending half of his working week in the classroom, it may make sense to visit some teachers 25 times during the academic year (e.g., beginning teachers or teachers who are working on explicit developmental problems and are seeking feedback and assistance from the principal), and to make only four or five visits with teachers who appear to be reasonably competent and successful.

I stated earlier that in many school districts principals or curriculum supervisors have had little, if any, formal training in using observation systems reliably or providing information to teachers in facilitating ways. However, in some school districts there is a growing willingness to make serious commitments to training principals and supervisory staff to play important

observational and instructional leadership roles in the schools. For example, the Little Rock (Arkansas) School District, under the guidance of Dr. Herb Williams and Dr. Carolyn Weddle is not only encouraging their elementary and secondary principals to spend one half of their time in classrooms, they are also providing training workshops and making materials available in order to develop principals' and supervisors' skills for collecting observational data. Such commitments are necessary if observational information is to be collected and used so that the information affects student achievement and attitudes.

Conclusions

In this paper, I have attempted to promote an interest in observational procedures. I feel that if observational data are collected in appropriate ways—with the intent of describing and understanding classrooms and stimulating possible actions—observational information can have positive effects upon classroom behavior. In particular, recent advances in research using more sophisticated observational procedures have begun to produce important results in terms of increased understanding of teacher behaviors that are associated with achievement.

I have advocated the use of observational techniques in inservice programs in schools. However, I have also tried to show needed restraint by pointing out that observational instruments are typically misused in school districts and have had little impact. There are some school districts which have advantageously used observational procedures to sensitize teachers to issues and to ways that help teachers to respond more creatively to particular classroom problems. Unfortunately, in too many school districts observational data are used as a symbol, having little substantive impact and value. If school districts are to realize the potential of observational research, it will be necessary to invest reasonable amounts of time and energy in understanding and developing techniques for using observational procedures.

Although I have alluded to particular research areas that I feel have useful concepts and potential value for observational work in school-based programs, I have resisted the opportunity to give my list of "best bets" about what behaviors one ought to observe in the classroom. I have shown restraint partly because there are few generic skills or concepts that appear to be important in a wide variety of teaching contexts. The contexts in which

data are being collected, and the purposes for which observation is being used, are important factors that have to be examined when developing a list of behaviors to code and a particular format for collecting information. The second reason that I have resisted providing a single set of questions that one ought to be concerned about, and have not recommended one or two observational systems that should be used, is because I feel that school districts ought to take a close look at some of the original research that has been conducted in the 1970s. Observational research has yielded a number of important concepts that can be used for examining and thinking about classroom behavior. As I stated earlier, school districts that are interested in using observational procedures for improving instruction can profitably begin to examine this literature and to consider its relevance in terms of their local goals and priorities.

If nothing else, observational research in the 1970s has convincingly demonstrated that teachers can and do make important differences in student learning. Furthermore, recent experimental research has also demonstrated that teachers can be trained to improve their instructional efforts. It seems to me that one very important way in which teachers' skills can be improved is through the successful use of classroom observation. Some schools are beginning to make more investments in training supervisors and encouraging the use of observational data in the classroom. However, *form* does not guarantee success, and only those school districts that are willing to invest in the careful study of teaching, and who are serious and careful about the applicative value of classroom observation, will be able to use it in a way in which classroom instruction is improved.

References

Amarel, M. *Literacy: The Personal Dimension.* A paper presented at the annual meeting of the American Educational Research Association, Los Angeles, April 1981.

Anderson, L.; C. Evertson; and J. Brophy. "An Experimental Study of Effective Teaching in First-Grade Reading Groups," *Elementary School Journal,* vol. 79, 1979, pp. 193-223.

Anderson, L.; C. Evertson; and E. Emmer. "Dimensions of Classroom Management Derived from Recent Research," *Journal of Curriculum Studies,* in press.

Bossert, S. "Task and Social Relationships in Classrooms: A Study of Classroom Organization and Its Consequences." American Sociological Association, *Arnold and Caroline Rose Monograph Series*. New York: Cambridge University Press, 1979.

Brophy, J. "Teacher Behavior and Its Effects," *Journal of Educational Psychology*, vol. 71, 1979, pp. 733-750.

_____. "Teacher Praise: A Functional Analysis," *Review of Educational Research*, vol. 51, 1981, pp. 5-32.

Brophy, J. and C. Evertson. *Learning from Teaching: A Developmental Perspective*. Boston: Allyn & Bacon, 1976.

Brophy, J. et al. *Student Characteristics and Teaching*. New York: Longman, Inc., 1981.

Brophy, J. and J. Putnam. "Classroom Management in the Elementary Grades: A Literature Review," in D. Duke, ed., NSSE 78th Yearbook, *Classroom Management*. Chicago: University of Chicago Press, 1979.

Chall, J. *Learning to Read: The Great Debate*. New York: McGraw-Hill, 1967.

Coleman, J. et al. *Equality of Educational Opportunity*. Washington, D.C.: Superintendent of Documents, U.S. Government Printing Office, 1966.

Cooper, H. and T. Good. *Pygmalion Grows Up: Studies in the Expectation Communication Process*. New York: Academic Press, in press.

Denham, C. and A. Lieberman, eds., *Time to Learn*. Washington, D.C.: U.S. Department of Education, 1980.

Doyle, W. "Classroom Tasks and Students' Abilities," in P. Peterson and H. Walberg, eds. *Research on Teaching: Concepts, Findings, and Implications*. Berkeley, Calif.: McCutchan Publishing Corporation, 1979.

Fisher, C. et al. *Teaching Behaviors, Academic Learning Time, and Student Achievement* (Final Report of Phase III-B, Beginning Teacher Evaluation Study). San Francisco: Far West Laboratory, 1978.

Good, T. "Teacher Effectiveness in the Elementary School: What We Know About It Now," *Journal of Teacher Education*, vol. 30, 1979, pp. 52-64.

_____. *Research on Teaching* (Tech. Rep. No. 207). Columbia, Mo: University of Missouri, Center for Research in Social Behavior, 1980.

Good, T. and J. Brophy. *Looking in Classrooms.* 1st ed. New York: Harper and Row, 1973.

Good, T. and J. Brophy. *Looking in Classrooms.* 2nd ed. New York: Harper and Row, 1978.

Good, T. and J. Brophy. "Changing Teacher and Student Behavior: An Empirical Investigation," *Journal of Educational Psychology,* vol. 66, 1974, pp. 390-405.

Good, T. and J. Brophy. *Educational Psychology: A Realistic Approach.* 2nd ed. New York: Holt, Rinehart and Winston, 1980.

Good, T. and D. Grouws. "Teaching Effects: A Process-Product Study in Fourth-Grade Mathematics Classrooms," *Journal of Teacher Education,* vol. 28, 1977, pp. 49-54.

Good, T. and D. Grouws. *Teaching Manual: Missouri Mathematics Effectiveness Project* (Tech. Rep. No. 132). Columbia, Mo.: University of Missouri, Center for Research in Social Behavior, 1978.

Good, T. and D. Grouws. "The Missouri Mathematics Effectiveness Project: An Experimental Study in Fourth-Grade Classrooms," *Journal of Educational Psychology,* vol. 71, 1979, pp. 355-362.

Goodlad, J. and M. Klein. *Looking Behind the Classroom Door.* 2nd ed. Worthington, Ohio: Charles A. Jones Publishing Co., 1974.

Hook, C. and B. Rosenshine. "Accuracy of Teacher Reports of Their Classroom Behavior," *Review of Educational Research,* vol. 49, 1979, pp. 1-12.

Jackson, P. and H. Lahaderne. "Inequalities of Teacher-Pupil Contacts," *Psychology in the Schools,* vol. 4, 1967, pp. 204-208.

Jackson, P. *Life in Classrooms.* New York: Holt, Rinehart and Winston, 1968.

Kounin, J. *Discipline and Group Management in Classrooms.* New York: Holt, Rinehart and Winston, 1970.

Martin, M. *Equal Opportunity in the Classroom* (ESEA, Title III: Session A Report). Los Angeles: County Superintendent of Schools, Division of Compensatory and Intergroup Programs, 1973.

Peterson, P. and H. Walberg, eds., *Research on Teaching: Concepts, Findings, and Implications*. Berkeley, Calif.: McCutchan Publishing Corporation, 1979.

Rohrkemper, M. "Teacher Self-Assessment," in D. Duke, ed., *That Learning and Teaching May Occur* (ASCD Volume on Classroom Management), in press.

Rosenshine, B. "Content, Time, and Direction Instruction," in P. Peterson and H. Walberg, eds., *Research on Teaching: Concepts, Findings, and Implications*. Berkeley, Calif.: McCutchan Publishing Corporation, 1979.

_____. "How Time is Spent in Elementary Classrooms," in C. Denham and A. Lieberman, eds., *Time to Learn*. Washington, D.C.: U.S. Department of Education, 1980.

Schwab, J. "The Practical: A Language for Curriculum," *School Review*, November 1969, pp. 1-23.

Stallings, J. *Learning to Look*. Belmont, Calif.: Wadsworth Publishing Co., 1977.

Stallings, J. "Allocated Academic Learning Time Revisited, or Beyond Time on Task," *Educational Researcher*, vol. 9, no. 11, 1980, pp. 11-16.

NSPER: 80

Phoenix Presentation

JOHN A. ZAHORIK
University of Wisconsin
at Milwaukee

Research on Teaching as a Basis for the Evaluation of Teaching

John A. Zahorik

To evaluate teaching it is essential that one know what good teaching is. For at least the past 50 years, education research has attempted to identify teaching that produces specific student achievements. The question addressed here is: Has research sufficiently identified "effective teaching" so that the generalized characteristics can be used as a basis for evaluating teaching?

Early research on teaching sought to identify effective practices by describing characteristics of successful teachers and by experimentally testing the effect of various teaching methods. Research on teacher characteristics consisted of asking students or experts to describe "good" teachers that they had known. These studies yielded characteristics such as enthusiasm, good judgment, and effective discipline. Subsequently, rating scales based on teacher characteristics were developed and used (Barr and Emans, 1930).

Teaching methods research consisted of comparing test results of groups of students who had been taught by different methods in either laboratory or naturalistic settings. Many methods were researched including the discussion method, the textbook method, the discovery method, the lecture method, and others.

Although teacher characteristics research and teaching methods research yielded practices that were believed to be effective, it became apparent that these types of research had limited value because of their generalness. Teacher characteristics research produced general teacher qualities that were difficult to translate into specific teaching acts, while teaching methods research focused on general patterns of teaching rather than on specific teaching acts. In addition, teacher characteristics research consisted of perceptions of effective teaching rather than demonstrated causes of student learning, and teaching methods research was plagued by confounding techniques as well as results.

Research of the last 20 years has attempted to overcome the problems of early research on teaching. It has focused on specific teacher behaviors such as questioning, praising, monitoring, and structuring. Some of this research emphasized

description, but most of the studies focused on both describing what teachers do in classrooms and determining the impact of instructional acts on student learning.

In the early phases of this work, research was conducted at many levels and in many content areas. Summaries of these studies as a whole (e.g., Dunkin and Biddle, 1974; Rosenshine and Furst, 1973) did not reveal teacher behaviors that were consistently related to student learning. Later, however, researchers began to concentrate on teaching at particular grade levels and in the same content areas. They concentrated on reading and mathematics instruction in the elementary grades, for example. Studies by Soar (1973), Stallings and Kaskowitz (1974), Brophy and Evertson (1974), Good and Grouws (1979), Solomon and Kendall (1976), and Bennett (1976) all have this focus and found the same result: A set of behaviors that Rosenshine (1976) has labeled "direct instruction" correlates positively and significantly with achievement of elementary school students in basic skills.

Direct instruction, according to Rosenshine, includes the following teacher behaviors:

1. Structures classroom time
2. Relies on texts and workbooks
3. Supervises and monitors student work
4. Directs activities without student choices or explanations
5. Asks questions and organizes learning around them
6. Uses narrow, single answer questions
7. Organizes subject matter to permit correct responses by students
8. Encourages students to attempt answers
9. Reinforces student responses
10. Gives feedback with an academic focus
11. After correct answer, asks new questions
12. After incorrect answer, gives correct answer

This type of teaching encourages high engagement levels of all students; it requires structured, accomplishable tasks; and it makes the teacher an active, vibrant teacher who is questioning, monitoring, and checking. Although on initial examination direct instruction might appear to be what traditional teachers have always done in classrooms, this is probably an inaccurate observation. In many traditional classrooms there may be teacher control of both students and subject matter, but there may also be tasks that are too difficult, a reluctance to monitor and supervise student work, interruptions in learning activities, infrequent feedback, and other behaviors that are contrary to direct instruction.

In contrast to direct instruction, Rosenshine found that the following teacher behaviors relate negatively to student achievement in reading and mathematics at the elementary level: 1) the use of games and interest centers, 2) permitting students to work independently, 3) encouraging student choice in activities, and 4) using open-ended questions.

Direct instruction, then, is a major finding, perhaps a milestone, in research on teaching. It correlates with achievement in basic skills; it consists of low-inference, easy-to-use behaviors; and it does not appear to harm student affect. It is not without weaknesses, however. First of all, it is correlational research rather than cause-effect research. Even though direct instruction has been found to be related to achievement gains in numerous studies, it is incorrect to say that direct instruction causes achievement gains. To make this claim, experimental studies are needed. Second, it uses achievement tests as a basis for judging effectiveness. Achievement tests are not capable of measuring all that was taught nor all that was learned in basic skills or other content areas. Third, this line of research carries a simplistic view of the relationship of teaching to learning. It does not adequately treat the variables that intervene between teaching and learning. Fourth, the relative significance of each component of direct instruction is not known. It is assumed that this composite of teaching acts can be applied by a teacher and the composite should be used in its entirety. Fifth, long-term effects of direct instruction in relation to basic skills and other learnings are not known.

Can direct instruction be used as a basis for the evaluation of basic skills teaching at the elementary level? Direct instruction is a promising concept and teachers who work in elementary schools with a disproportionate number of students with low achievement in basic skills may want to use it, but it would be premature to *require* teachers to use it. There are too many questions that need to be answered about its nature, functioning, and effects before it can serve as a basis for evaluating teaching. We do not know, for example, what the effects of direct instruction would be when applied to the teaching of social studies, chemistry, auto mechanics, or other subjects in high school.

We can conclude that research has not, at least up to this time, identified effective ways of teaching on which the evaluation of teaching can be based. Whether research will be productive in the future invites speculation. It is clear, however, that if research is to be productive, attention will have to be given to

experimental studies and studies that have a broader, holistic view of the complex classroom. Indeed, some argue that the classroom is too complex to ever be understood, and that even if effective characteristics of teaching are identified, classrooms will remain so unique that the transferability of general behaviors is dubious.

Both teaching and the evaluation of teaching would be simple if we had a scientifically verified set of effective teacher behaviors. The truth is, we do not. Teachers and supervisors will have to look elsewhere for the grounding principals of their profession.

References

Barr, A. S. and L. M. Emans. "What Qualities are Prerequisites to Success in Teaching?" *Nation's Schools,* vol. 6, 1930, pp. 60-64.

Bennett, N. *Teaching Styles and Pupil Progress.* Cambridge, Mass.: Harvard University Press, 1976.

Brophy, J. E. and C. M. Evertson. *Process-Product Correlations in the Texas Teacher Effectiveness Study: Final Report.* Austin, Tex.: University of Texas, 1974.

Dunkin, M. J. and B. J. Biddle. *The Study of Teaching.* New York: Holt, Rinehart and Winston, 1974.

Evertson, C. M., C. E. Anderson, L. M. Anderson and J. E. Brophy. "Relationships Between Classroom and Student Outcomes in Junior High Mathematics and English Classes," *American Educational Research Journal,* vol. 17, 1980, pp. 43-60.

Good, T. L. and D. A. Grouws. "Teaching and Mathematics Learning," *Educational Leadership,* vol. 37, 1979, pp. 39-45.

Rosenshine, B. "Classroom Instruction," in N. L. Gage, ed., NSSE 75th Yearbook, *Psychology of Teaching Methods.* Chicago: University of Chicago Press, 1976.

Rosenshine, B. and N. F. Furst. "The Use of Direct Observation to Study Teaching," in R. Travers, ed., *Second Handbook of Research on Teaching.* Chicago: Rand McNally, 1973.

Soar, R. S. *Follow Through Classroom Process Measurement and Pupil Growth (1970-71): Final Report.* Gainsville, Fla.: University of Florida, 1973.

Solomon, D. and A. J. Kendall. "Individual Characteristics and Children's Performance in 'Open' and 'Traditional' Classroom Settings," *Journal of Educational Psychology*, vol. 68, 1976, pp. 613, 625.

Stallings, J. A. and D. H. Kaskowitz. *Follow Through Classroom Observation Evaluation, 1971-73*. Menlo Park, Calif.: Stanford Research Institute, 1974.

Qualitative Versus Quantitative Approaches to Observing Teaching

John A. Zahorik

Typically, there are two purposes for the evaluation of teaching: instructional and administrative. Teachers are evaluated in an effort to improve their teaching and, subsequently, students' learning, and they are evaluated to facilitate administrative decisions about teacher retention, assignment, and merit pay. Observation of teaching is important for gathering data for both of these purposes, but it is more important for instructional purposes than it is for administrative purposes. Many times information that is helpful in making administrative decisions comes to the attention of administrators without classroom observation. To improve teaching, however, classroom observation is essential.

Supervisors employ observation in a variety of ways when evaluating teachers. Some schedule and complete formal observations, while others make spontaneous observations. Some simply watch the teacher's classroom behavior as unobtrusively as possible; others take part in classroom activities by tutoring students or actually teaching the class for awhile. Some use observation instruments, while some only make mental notes.

Most school districts or education institutions have a set of criteria that they use to evaluate teaching. Frequently these criteria are expressed in the form of observation instruments or rating scales that are to be completed during observation, although they could be in the form of a rating scale that is completed after a range of information is assembled including data from observation. An example of one such observation instrument that districts often use is the Flanders (1960) instrument. With this instrument the supervisor can record the frequency and pattern of 10 verbal behaviors: 1) accepting feelings, 2) praising, 3) accepting ideas, 4) questioning, 5) lecturing, 6) directing, 7) criticizing, 8) student responding, 9) student initiating, and 10) silence or confusion.

Flanders's instrument is a research instrument, and the practice of actually observing teaching is largely an idea that comes from research. Certainly, the practices that supervisors use originate from many sources, but there appears to be a distinct relationship between the *type* of supervision in schools and the

type of research carried on by education researchers. That is, supervisors apparently fashion their roles from their understanding of education research. At present many researchers in education are turning to what amounts to a new approach for them—namely, ethnography. If what I said above is true—that supervisors and evaluators fashion their roles from education research—we can expect them to be influenced by the ethnographic approach, which is a qualitative approach. The hope is that this influence will improve supervision and evaluation and, thus, teaching and learning.

Ethnography, which has its roots in anthropology, is based on two perspectives of human behavior—the ecological and the phenomenological (Wilson, 1977). The thesis of the ecological perspective is that behavior is influenced by the setting in which it occurs. It is influenced by the physical arrangements and by the beliefs and expectations of participants in the setting. The thesis of the phenomenological perspective is that behavior can only be understood if it is examined from the subject's point of view. From this perspective, reality is what people perceive it to be. Taken together and applied to research on teaching, these perspectives mean that researchers must be concerned with the whole classroom setting and with teachers' thinking if they are to begin to understand the teaching-learning situation.

In contrast to the qualitative focus in ethnography, the dominant method in social/behavioral research over the past 50 years has emphasized quantitative analysis. This research method is patterned after the basic methods of research in the physical sciences and is based on a positivistic view of human behavior (Bogdan and Taylor, 1975). From this perspective human behavior can be fragmented and studied with little regard for individuals' perceptions. Reality exists apart from people. Applied to research on teaching, this conception of human behavior means that isolated classroom events and their relationships will be studied and teachers' and students' perceptions will be neglected.

In addition to, and because of, differences in perspectives concerning human behavior between those two types of research, differences also exist in the types of questions that are asked, types of methods that are used, and types of findings that are reported.

Hinely and Ponder (1979) offer a useful distinction between the research questions that quantitative researchers ask and those that qualitative researchers ask. Quantitative researchers, or what Hinely and Ponder label as *improvers,* seek to answer:

How can things be changed? Qualitative researchers, or what they label as *describers,* seek to answer: What is happening here; why is it happening; and what does it mean in the classroom context? Quantitative research, then, seeks to verify practices while qualitative research seeks to understand classroom life.

In relation to research methods, quantitative research is concerned with variables, treatments, sample size, randomization, and statistical description and inference. Data are collected with tests, questionnaires, or checksheets. In contrast, the qualitative researcher is concerned with choosing a site, gaining site access, researcher roles and rapport, keeping field notes, and developing typologies. Data are collected through participant and other types of observation, interviews, and personal documents.

As would be expected, the outcomes of these two types of research differ dramatically. Quantitative research identifies practices that have been demonstrated to be effective. It is able to do this because it isolates and insures variation in the independent variable being studied; it uses reliable and consistent techniques to measure both the independent and the dependent variables; and it obtains data from a representative random sample sufficiently large to permit generalization. If this type of research does not produce cause-effect data that can guide teaching practice, it produces correlations between teaching practices and measures of student growth that often are presented in a cause-effect format and which guide practice. Most teaching methods and teacher behavior research exemplify this kind of outcome.

The other type of research, qualitative, results in descriptions, understandings, and insights. It cannot identify practices that have been demonstrated to be effective, because it examines the whole setting rather than isolated and highly controlled independent and dependent variables. Qualitative research uses measurement techniques that are subjective, techniques that vary from context to context. It also treats small samples in depth rather than large samples in breadth. This is not to say that causes and effects are not aspects of qualitative research. Causes and effects are implicit in the terms "understanding" and "insight." These causes and effects, however, deal with many aspects of classroom life, not just impact on student learning, and they are tentative hypotheses rather than confirmed practices.

Flanders's checklist and other similar supervisory instruments are mainly quantitative instruments. They provide data about

the teaching act, but they do not provide a holistic account of the classroom. The seven teacher behaviors and two student behaviors that the instrument records are only a partial list of the behaviors that occur in the classroom. Also, the instrument provides no data about the teacher's values; the teacher's objectives for the particular class session; students' past experiences, interests, and abilities; the nature of the classroom setting; types of materials available; and numerous other variables.

Qualitative research is time consuming and requires special skills on the part of the researcher. A supervisor cannot be expected to become a qualitative researcher, but there are several actions that the supervisor could take to make supervision more qualitative, and possibly, more effective in improving teaching. The following are suggested:

1. Attempt to understand what is observed in the classroom from the teacher's point of view. This means that the supervisor will need to know the teacher's values, purposes, and past experiences.

2. Focus on the whole classroom setting rather than a few isolated teacher behaviors. This means that the supervisor will need to study the nature of the students and the total classroom setting.

3. Whenever time permits, be a participant observer. Teach a small group or the total class. Many insights will emerge that would not become known from nonparticipant observation.

4. Although some observations need to be made with predetermined instruments, make some observations in which categories are permitted to emerge. Attempt to see what is there before classifying what is there.

5. Interpret as you observe. Attempt to infer meaning of an action as it occurs rather than separate judgment from description.

6. Spend as much time observing as possible.

7. In addition to observing, discuss observations and impressions with teachers. Also, probe to help teachers clarify their values and strive for values-behavior consistency.

8. Make your values and prejudices known, but do not force them on teachers. Discussion of honest differences in beliefs can be profitable for supervisors and teachers.

9. Treat whatever conclusions are drawn from observation as hypotheses to be tried and tested rather than as terminal truths.

Observation of teaching is a difficult task. The addition of qualitative techniques does not make it less difficult. It may, however, make it better. It may make an important difference in the improvement of teaching.

References

Bodgan, R. and S.J. Taylor. *Introduction to Qualitative Research Methods.* New York: John Wiley, 1975.

Flanders, N.A. *Teacher Influence, Pupil Attitudes, and Achievement.* Final Report, Cooperative Research Project no. 397, U.S. Office of Education. Minneapolis, Minn.: University of Minnesota, 1960.

Hinely, R. and G. Ponder. "Theory, Practice, and Classroom Research." *Theory Into Practice,* vol. 18, 1979, pp. 135-137.

Wilson, S. "The Use of Ethnographic Techniques in Educational Research," *Review of Educational Research,* vol. 47, 1977, pp. 245-265.

Values as a Basis for the Evaluation of Teaching

John A. Zahorik

Research on teaching has not been successful in identifying effective ways of teaching. It may be productive in the future, but neither teachers nor supervisors can wait for the guidance that research might provide. Teaching demands action now. Committed supervisors are attempting to improve teaching now. This paper examines and prescribes values as an appropriate basis for teaching and for the evaluation of teaching.

Besides data from research on teaching, there are other possible bases for the evaluation of teaching. Among the alternatives are: 1) following the prescriptions of experts, 2) modeling successful teachers, 3) adhering to a philosophical position, and 4) following a set of values. Unlike research, any one of these alternatives would be sufficiently predictive to be used at present. I will take the position, however, that all of these alternatives are inappropriate except values.

There is no shortage of prescriptions by experts concerning what teachers ought to do in classrooms. In the area of classroom management, for example, detailed prescriptions are readily available. A teacher could decide to employ the behaviors recommended by Glasser (1969) or Gordon (1974), or the teacher could justify a set of behaviors that he or she normally uses by showing their consistency with Glasser's or Gordon's ideas. Using an expert's pronouncements as a basis for teaching is probably a widespread practice. But, on what are the expert's prescriptions based, and how does one decide which expert's ideas to accept and follow? The expert's suggestion could be based on research but, as we have seen, this is improbable. It is more likely that the prescriptions flow from a value position that the expert holds. Teachers cannot randomly choose experts to follow. If they do, inconsistent and contradictory behavior may result. If an expert's prescriptions are to be used as a basis, teachers must decide first if the expert's values are consistent with their own values. This requires the teachers to determine what their values are. It makes teacher values the primary basis for decision making.

In the process of being a student, in learning to become a teacher, and in teaching on a full-time basis, one comes into

contact with many teachers. All of these teachers are models of teaching. Some are models of good teaching, and some are models of poor teaching. A teacher can decide to use a behavior that another teacher uses and justify it on the basis that the other teacher is a competent, effective teacher. The decision to choose one teacher to model and not another cannot be an unconscious decision, however. If it is, the teaching behavior borrowed from the model may again be inconsistent with and contradictory to other teaching behaviors that the teacher uses. The decision about whom to model has to be made on the perceived congruity between the model's values and the teacher's values. To determine whom to model, then, requires the teacher to determine what his or her values are. Again, the teacher's values are the primary basis for deciding how to teach.

Philosophical position is also an unsuitable basis for teaching unless teachers first clarify their values. One simply cannot choose a philosophical position such as existentialism, essentialism, or experimentalism and expect sincere, effective teaching behavior to result. One must embrace the philosophical position, and this requires an understanding of one's values.

Expert prescriptions, modeling of others, and philosophical position as useful bases for deciding about which teaching behaviors to use depends on teachers knowing their values. This is also true for research on teaching. We assume that if research demonstrates that one behavior is more effective in bringing about student learning than another, teachers will decide to use the behavior in their teachings. The assumption is that teachers are valueless technicians who will use any and all behaviors that research certifies. This action on the part of teachers is highly unlikely, even if research is able to identify behaviors that facilitate student learning. Teachers and others have values, and research findings get sifted through value systems before they influence behavior. How else can one explain the continued aversion to exercise, the persistence in cigarette smoking, and the refusal of many to wear seatbelts in their cars? The research evidence in each of these areas is substantial, yet behavior is not changed appreciably.

Values, then, are a major basis for deciding how to teach and probably serve as the basis for the evaluation of teaching. When teachers deliberate about which teaching behaviors to use, they most likely look to their values. Although it may be dissatisfying to many, values are the most helpful and powerful guideline available to teachers at present.

Good teaching, from the perspective of values, is teaching which is consistent with a well-developed set of values. A well-developed set of values is one that is consciously considered and carefully examined; results in a noncontradictory, meaningful whole; and is moderately flexible. Teaching that is consistent with values will be purposeful, conscious, and controlled. That is, the teacher has reasons for using a particular behavior; the teacher is aware of the behaviors that are and are not being used; and the behavior is only used to the extent deemed appropriate by the teacher.

Poor teaching, from the perspective of values, is teaching that emerges from an inadequate set of values or that is inconsistent with the teacher's set of values. A set of values that is whimsical and is not a product of deliberation, is internally contradictory, and is either too rigid or too flexible will not result in good teaching. Teaching that contradicts values is also poor teaching, because it has no hope for being a consistent whole. If it is not consistent and does not result in a meaningful whole, the teacher's values may be superficial and need to be reformulated.

If sufficient research were available to use as a basis for evaluating teaching, the supervisor's task would be relatively simple. The supervisor would only need to observe teachers in their classrooms and compare their behaviors with those that research has found to be effective. With values as the basis, the supervisory task is more difficult. The supervisor must come to know a teacher's values as well as his or her behavior. In addition, the supervisor must become aware of his or her own values as well as the values manifest in the classroom and school setting.

Developing a set of values is a long, if not endless, process. Coming to know one's beliefs and convictions about what is good, desirable, or preferable is difficult. It may be helpful in developing a value statement to focus on four elements of education: students, knowledge, teacher, and aims. Asking and answering appropriate questions about each of these elements could help to focus a set of values for teachers.

The major question that needs to be answered about students is: What are students like? To answer this question teachers would have to decide what they believed about a series of other questions. Are students capable of making responsible decisions? Are students self-motivated? Do students naturally engage in exploratory behavior? Can maturation be influenced? Is student choice an important aspect of learning?

The major question that needs to be answered about the element of knowledge is: How should knowledge be treated? This question is also composed of a series of subordinate questions. What substance, skill, or affective learnings are essential for all students? What are the functions of subject matter? How should subject matter be organized? Is subject matter personal?

The major question that one would have to answer about the teacher is: What role should the teacher assume in the classroom? Again, this question subsumes a number of other questions. Should the teacher use a behavior modification style of classroom management or should some other style be used? Should the teacher be a knowledge source, a facilitator of learning, or a diagnostic technician? Should the teacher help students clarify their own values or instill accepted values of society?

The major question to be ansered about aims is: What should the major aim of education be? More specific questions are: Should the aim be academic growth in basic skills? Should the aim be a healthy self-concept? Should the aim be social growth? Should the aim be economic security?

Answers to questions such as these could come from many sources including research, experts, philosophy, other teachers, and personal teaching experience.

When the principles of teaching, and the principles of evaluation as well, are grounded in a set of values, the supervisor's objectives will be to increase consistency in the value-based teaching behavior and to promote values clarification. Observation is essential in the former while challenging conversation is necessary for the latter. The complete process requires an open-minded supervisor who knows what his or her values are, but does not force them on others, and who believes that there are a variety of effective ways to teach.

References

Glasser, W. *Schools Without Failure*. New York: Harper and Row, 1959.

Gordon, T. *Teacher Effectiveness Training*. New York: Peter H. Wynden, 1974.

NSPER: 80

Toledo Presentation

LARRY A. BRASKAMP
University of Illinois
at Urbana-Champaign

What Research Says About the Components of Teaching

Larry A. Braskamp

Before considering what research says about the components of teaching, I want to first reflect a little upon what is meant by teaching and specifically what is meant by effective teaching. In assessing teaching, we are not just interested in describing all the various activities that go on in teaching, but we would like to be able to delineate more closely what we mean by effective instruction and teaching.

The history of the research on teaching is long and varied. Perhaps the best book on this topic is the one edited by Penelope Peterson and Herbert Walberg (1980) titled *Research on Teaching: Concepts, Findings, and Implications.* The editors use a temporal theme in organizing the chapters—the history of teaching, its current status, and its projected future. Donald Medley, in the first chapter, divides the history of research into four phases. In the first phase, teaching effectiveness is assumed to be the result of the teacher's personality characteristics or traits. In the second phase, effectiveness is viewed as a consequence of teaching methods (comparative experiments were conducted to test the relative effectiveness of various teaching methods such as student versus teacher-centered classrooms). The third phase, predominant in the 1960s, included the systematic counting of specific teacher and student behaviors in an attempt to associate teacher behavior with student outcomes. In this "process-product research," teaching styles or classroom climate were studied rather than teacher characteristics. Interaction analysis was a prominent strategy during this phase. In the most recent phase, the mastery and employment of specific teaching competencies have served as the cornerstone for studying and understanding effective teaching. The investigation of teaching competencies includes an examination of "not only how teachers behave but when and why they behave as they do" (Peterson and Walberg, 1980, p. 16).

Rosenshine (1978) discusses a fifth approach in which the focus is on what the students do in the classroom. Direct instruction and the time students spend engaged in mastering academic skills are the major variables of study.

In each of these five major approaches to the study of instruction, different definitions of effective teaching are used and dif-

ferent components of teaching are included in the study of teaching. The history of research on teaching illustrates the close linkages between methodological approaches to the study of teaching and the theories of instruction currently in vogue. For example, today there is a considerable interest in information processing, i.e., how do students perceive how to learn from teaching. Thus the new developments have to do with interest in student variables and the study of cognition and individual differences in the way students process information.

In this symposium we are concerned with the evaluation of teaching in which the evidence or evaluative information is collected by some form of structured observation. However, this restriction does not allow us to ignore the conceptual framework or theoretical orientation we employ in selecting variables to study and in interpreting the evaluative informaion. In short, how we decide to evaluate instruction depends upon the conceptual framework we wish to use in understanding and trying to analyze precisely what we mean by effective instruction.

Four Guidelines

Next I wish to present four guidelines that I use in evaluating teaching. These guides are partially based on the research on instruction, partly on methodological problems in assessing evaluation, and partly on my own personal views or values of effective instruction. I hope each of these will become more evident as we focus on evaluation in these two days.

1. Since there is no one standard in which effective teaching can be evaluated, diversity in teaching styles is to be advocated. This position is based partly on the conflicting results that are found in the research on instruction. I think we are more prudent in assuming that teaching is as much of an art as it is the mechanical application of a series of laws of learning. Thus it is not advisable to be highly prescriptive in our advice on effective instruction. Cronback (1975), in his analysis of the two disciplines within psychology, argues that generalizations should come late if at all in our research. Instead we must take into account local factors and local traditions; we need to emphasize local description. When we do this, any generalization is a working hypothesis, not a conclusion. From the reviews of research on evaluation of college teaching, e.g., Kulik and McKeachie (1975), Centra (1980), we know that people disagree on what constitutes effective instruction and that different types of evidence such as student

ratings, colleague ratings, and student achievement do not always point with consensus to the most effective teachers.

2. The definition of teaching effectiveness is contingent upon the expectations and instructional goals of the institution, the type of students enrolled, course goals, and the instructor's teaching style. McKeachie (1976), in his American Psychological Association Presidential Address, lists nine characteristics of effective teaching that he says are based on what psychologists have learned about human behavior. One such characteristic is *flexibility and variety*. Teachers need to be flexible and employ a variety of approaches because of "interactions among student characteristics, teacher characteristics, goals, subject matter, and methods..." (p. 826).

3. In evaluating instruction, a set of criteria is used, at least implicitly. When we evaluate something, we are employing a set of criteria in its evaluation; i.e., we use a set of dimensions or characteristics or bases called criteria which we are willing to defend as appropriate for assessing the value or worth of that something. For example, if we are evaluating a car we use criteria such as miles-per-gallon, style, riding comfort, or safety in assessing its worth or value. The same process is involved when we evaluate an instructor or a particular course. Operationally we are selecting dimensions which we think are the most appropriate and the most defensible to assess the effectiveness of a course. There are a host of criteria, many of which are implicit or for which we have difficulty in finding behaviorial indicators. For example, when we think about evaluating a course such as psychology 101, one of the more obvious criteria we could use is the amount of student learning—e.g., the number of technical terms that freshmen know at the end of the semester, or their ability to compare different forms of psychotherapy or compute a chi square test. Other possible criteria can be student interest in human behavior or greater self-understanding. The possible list of criteria can go on and on.

So far we have not mentioned anything about the ability of the instructor to communicate effectively, the ability of the instructor to motivate the students, or any other characteristic of the instructor. We also have not mentioned anything about the sequence of the course, the difficulty level of the course, the clarity of objectives of the course, or the friendliness of the instructor—some of the more obvious salient features we often use in evaluating a course. Thus in evaluating anything, we should first ask: What criteria were employed in evaluating that particular course? How are those criteria selected? How can they be

defended? The next question is: Can appropriate indicators of these criteria be employed? The answer to this question is basically one of measurement. Does the evidence adequately reflect the criteria to be used? Is the evidence a good indicator of the criteria? For example, just because we decide that student learning ought to be a criterion, we still have to go the next step and determine what measurement procedures we are going to employ to collect the evidence or evaluative information. The third step involves the selection of standards; i.e., what level of performance (either absolute or relative) is needed before one can make judgments about value. Thus we have a sequence of steps that we have to go through in the practical evaluation of a course. We have to determine the criteria; we have to find indicators, which means locating reliable and valid measures; and then we have to determine standards of acceptable performance for each criterion before we are able to assess any kind of value or worth. Again, we can use student achievement as an example. We have to find a test; we have to collect evidence; then we have to determine the level of performance that is considered to be indicative of good instruction. The standards that we employ could be absolute; that is, we could make an *a priori* judgment about acceptable performance; or we may employ a relative standard, such as comparing one class with another class or comparing the class with previous classes.

So far we have said nothing about the goals and objectives of the course. One can select a set of criteria which may or may not be related to the goals and objectives of the course. The selection of criteria will probably depend upon one's philosophy toward evaluation as well as one's definition of instruction. It is commonly held that a major criterion in the evaluation of instruction is to assess the extent to which the students have obtained the objectives of the course. Thus teaching effectiveness can be defined as the amount of progress students make on teacher-defined goals (Aubrecht, 1979). Others tend to evaluate a course based not upon what students learn but rather upon the transactions or behaviors of the teacher in designing a course. Still others advocate a set of criteria selected by some external group of consumers and may include more than what is implied by the goals and objectives of the course. In short, criteria are implicitly used. They may or may not reflect the goals of the course, but practically speaking, one cannot avoid selection of criteria in evaluation.

4. Structured observation (observation scales, checklists, rating scales) is one measurement procedure that can be

employed in the evaluation of instruction. Instruction is too dynamic and too complex to be evaluated by only one measurement method. That is, we need to examine instruction from a variety of perspectives in which multiple techniques of data collected from a number of sources are used. As we shall discuss more thoroughly later on, not all audiences will view the quality of instruction similarly, and not all data techniques provide the same type of information, although the convergent validity among sources (teachers, peers, self, colleagues, and students) and methods of measurement (student ratings on fixed alternative items, written comments, and student comments in group interviews of college teachers) is encouraging.

Research on College Teaching

Because of my interest and experience in teaching at the postsecondary level, my review of the research literature will be limited to those studies that have been conducted at the collegiate level. Unfortunately, most of these studies are still limited to a consideration of the personality characteristics of effective instructors. The one major exception is a series of studies on the effectiveness of personalized systems of instruction, e.g., the Keller Plan (Kulik and Kulik, 1979).

There are two major reasons why teacher personality has been the focus. First, college professors view themselves as models and purveyors of knowledge, and they are a very important factor in the teaching/learning process. There is a prevalent belief at the collegiate level that the instructor is the person who makes the difference, rather than the curriculum, the structure, or the host of audiovisual techniques ranging all the way from overheads to complex PLATO computer systems. Many in higher education still doggedly adhere to the dominance of that unique faculty member who, through his style and modeling behavior, will not just immediately influence students but will have an enduring and lasting effect on how students behave as problem solvers and as future scholars and practitioners. Second, those studying instruction have been more empirical than theoretical in their orientation. Descriptive data analyses are more common than theory testing. Consequently, many empirical investigations have focused on the personality characteristics of instructors and course particulars. Thus I will not give a detailed review of these studies here, because others have already written extensively about them (Doyle, 1975; Centra, 1980).

For our purposes we can organize the research on teaching around four questions:

1. What are the components or characteristics of instruction at the collegiate level that have been identified by factor-analytic studies on student ratings of instruction?
2. What are the characteristics of instruction in teaching based upon written comments by students?
3. What are the characteristics of the "ideal" college teacher (characteristics of superior quality teaching) as seen by students and professors themselves?
4. What factors do colleagues consider when they evaluate instruction?

I have listed two syntheses of factor-analytic studies that have been conducted on student ratings of instruction. The first is a listing of factor-analytic studies reviewed by Brandenburg, Derry, and Hengstler (1977) as background material for the development of our student rating system at the University of Illinois. The factors obtained in 16 factor-analytic studies are presented in Table 1. On the basis of the review, the authors concluded that instructor skills, student rapport, teaching structure, course difficulty, and student outcomes are the five major dimensions that are included in most rating scales which are used to evaluate instruction. This list includes one characteristic about the instructor and his or her ability to communicate, one about student/teacher interaction and ability to form good relationships, another about the structure and organization of the course. The remaining two are about student assessment of the difficulty of the class and self-reports on how much students have learned, either cognitively or affectively.

The second set of factors as shown in Table 2 are those developed by John Centra in his book, *Determining Faculty Effectiveness.* He lists seven factors or categories of ratings that are included in most student rating forms. If we examine these we will see that the same factors are included here as in the previous synthesis and that grading and examinations, as well as the global overall ratings of effectiveness, are included as the additional two factors. Grading is a factor that is an important area to be evaluated by most instructors (more of this in the later talk on values). The inclusion of overall ratings as a factor is not so much the direct result of factor-analytic studies as it is the consequence of the relationship between various dimensions of factors and relationships with an overall assessment of instruction (more will be said about this later in the section on validity).

Table 1
Factors Obtained from Sixteen Factor-Analytic Studies Based on Student Ratings of Collegiate Instruction[1]

FACTORS IDENTIFIED / STUDIES	Barnoski & Sockloff 1976	Despande, et al. 1970	Doyle 1977	Doyle & Whitely 1974	Frey, et al. 1975	Gibb 1955	Greenwood, et al. 1974	Hartley & Hogan 1972
Skill	Lecture Presentation Effectiveness	Stimulation (Communication)	Expositional Clarity (See Structure)	Expositional Skills	Clarity of Presentation	Communication	Facilitation of Learning Voice Communication	Overall Evaluation
Rapport	Professor-Student Relationship	Affective Merit	Attitude Toward Students	Attitude Toward Students	Personal Attention	Friendly-Democratic	Rapport Openness	Student-Teacher Interaction
Structure		Cognitive Merit	Expositional Clarity (See Skill)		Organization/ Planning	Organization	Obsolescence of Presentation Currency of Knowledge	Structure/ Organization
Difficulty		Stress			Work Load	Academic Emphasis	Evaluation	Load or Difficulty
Outcomes			Outcome Motivation	Motivation of Interest in Subject Matter	Student Accomplishment			General Cognitive Development Field Specific Dev. Relevance
Other	Text & Readings Examination Effectiveness		Stimulation of Student	Stimulation of Thinking Generalization of Course Content	Class Discussion Grading	Residual	Commitment to Teaching	

	Hildebrand et al. 1971	Holmes 1971	Isaacson et al. 1964	Linn, Centra, Tucker 1975	Pohlmann 1972	SIRS MSU 1971	Solomon 1966	Umble & Whitten 1977
Skill	Dynamism/ Enthusiasm	Instructor Presentation	Skill	Lectures	Clarity of Communication	Instructor Involvement	Energy vs. Lethargy	Professional Competency
Rapport	Instructor-Student Individual Interaction	Instructor Evaluation	Rapport	Teacher-Student Interaction		Student-Instructor Interaction	Lecturing vs. Participation Warmth vs. Coldness Criticism vs. Tolerance	Receptivity
Structure	Organization/ Clarity	Student Stimulation*	Structure	Course Objectives and Organization	Class Management	Course Organization	Control vs. Permissiveness	
Difficulty			Overload	Course Difficulty and Workload	Difficulty	Course Demands		Potency-Difficulty
Outcome		Student Stimulation *			General Attitude	Student Interest		Motivation
Other	Analytic/ Synthetic Approach Instructor-Group Interaction		Group Interaction Feedback	Exams Readings and Assignments	Dependability Course Overlap Textbook		5 others	Personality Professional Behavior

[1] D. C. Brandenburg, S. Derry, and D. D. Hengstler, "Validation of an Item Classification Scheme for a Student Rating Item Catalog," a paper presented at the National Council on Measurement in Education annual meeting, Toronto, April 1978.

Table 2

Categories of Ratings and Examples of Items*

1. **Organization, Structure, or Clarity**
   - •Material presented in an orderly manner
   - •Instructor well prepared for each class
   - •Class time well spent
   - •Course well organized
   - •Instructor made clear what we were expected to learn
   - •Considerable agreement between announced objectives and what was taught

2. **Teacher-Student Interaction or Rapport**
   - •Instructor readily available for consultation with students
   - •Instructor seemed to know when students didn't understand the material
   - •Instructor actively helpful when students had difficulty
   - •Students felt free to ask questions or express opinions
   - •Instructor seemed concerned with whether students learned the material

3. **Teaching Skill, Communication, or Lecturing Ability**
   - •Instructor used examples or illustrations to clarify the material
   - •Instructor spoke audibly and clearly
   - •Instructor presented material clearly
   - •Instructor summarized or emphasized major points in lectures or discussions

4. **Workload, Course Difficulty**
   - •In relation to other courses, this workload was heavy
   - •Instructor tried to cover too much material
   - •Reading assignments were very difficult
   - •Course challenged me intellectually
   - •I put a great deal of effort into this course

5. **Grading, Examinations**
   - •Instructor told students how they would be evaluated
   - •Examinations reflected the important aspects of the course
   - •Instructor made helpful comments on papers or exams
   - •Instructor assigned grades fairly and impartially

6. **Impact on Students, Student Self-Rated Accomplishments**
   - •I learned a great deal in this course
   - •This course generally fulfilled my goals
   - •This course stimulated me to want to take more work in the same or a related area

7. **Global, Overall Ratings**
   - •Instructor's effectiveness as a teacher was: (excellent to poor)
   - •Overall value of the course was: (excellent to poor)
   - •Instructor made a major contribution to the value of this course
   - •General quality of lectures was: (excellent to poor)
   - •General quality of class discussions was: (excellent to poor)

---

[1] This listing is taken from J.A. Centra's *Determining Faculty Effectiveness*, San Francisco: Jossey-Bass, 1980. Used by permission.

It should be pointed out that the inclusion of these factors is not normative; the factors are descriptive. That is, they indicate the clustering of items that are included in the student ratings of instruction. However, they do have a normative base if we examine the history of the development of these surveys. Early on faculty and students were asked to write down or select characteristics of a good or effective teacher (e.g., Hildebrand, Wilson, and Dienst, 1971). Based upon these descriptions, student rating forms were developed, factor analyzed, and revised. Thus the original item pool is based upon studies that have been done in which effective instruction has been the reference point for establishing a pool of items.

A second basis for identifying characteristics of instruction is through a content analysis of written comments by students. In a study conducted in our office we read student comments on 60 instructors representing all professorial ranks and 10 different content areas (Braskamp, Ory, and Pieper, 1981). Based upon a content analysis of student comments to four open-ended questions ("What are the major strengths and weaknesses of the instructor?" "What aspects of this course were most beneficial to you?" "What do you suggest to improve this course?" "Comment on the grading procedures and the exams."), 22 different categories were developed. Table 3 lists the categories and key words which were used to define these categories. This list closely parallels the list based on the factor-analytic studies. There are more categories that deal with the ability of the instructor to communicate clearly, the ability to stimulate students, and the instructor's own attitudes towards the course. One of every two comments was about a characteristic of the instructor, although two of the four questions related to the course. Nearly one-fourth of all the comments about the instructor were about his or her pedagogical abilities, whereas the second most common characteristic was instructor knowledge of the field followed by instructor's personal relationship with students (e.g., instructor as a person, degree of rapport, and helpfulness). Over one-third of the student comments were about the course content and materials. Assignments and workload accounted for an additional 40% of all comments. Organization of the course—i.e., pace, integration, clarity of objectives—was commented on less than 9% of the time. Thus students tend to focus on instructor characteristics and what they have learned but do not frequently critique the structure and organization of the course.

Third, quality of instruction has been examined through both student and colleague identification of characteristics they

Table 3

Classification Scheme Based on Student Written Comments about Instructor, Course, and Grading*

Instructor

1. Attitude—interest in, enthusiasm for, and attitude toward teaching, course and subject matter.
2. Rapport—concern and care for individual students; warmth, understanding; knows student names; bias toward a few students.
3. Person—professional, nervous, impatient, amiable personality; has annoying mannerisms; sincere; expresses humor, self confident.
4. Knowledgeable—knows subject matter, material; well versed.
5. Stimulation—challenging, fascinating; creates interest in field, stimulates thought; encourages new ideas; boring.
6. Rigor—expects too much, too little; challenges class to improve; poor discipline in class.
7. Openness—flexible; open to differing and opposite ideas; self righteous.
8. Ability—makes clear and concise presentations; talks too fast, too slow; easy to take notes; explains well; uses good examples.
9. Preparedness—follows an outline; poorly prepared for class; good organization of lecture.
10. Helpfulness—patient; helpful; readily available; spends time with individual students outside class.
11. Teacher overall—excellent teacher, good, poor role model; desire to take more courses.
12. Improvement—instructor became more relaxed, better lecturer at end of semester.
13. Miscellaneous—noncodable responses.

(continued)

Course

14. Organization—clarity of expectations, objectives; pace was even, uneven; good, poor course structure; lectures and labs were integrated.
15. Content—learned subject matter; learned writing, speaking, calculating, selling skills.
16. Materials—textbooks, handouts, audiovisual materials were good, poor, used effectively; lab equipment.
17. Requirements—too many, too few required assignments; appropriate projects; vague assignments.
18. Lecture—lectures are good, poor, not worth attending, integrative.
19. Discussion—class interactions are good, relevant; discussions are interesting, superficial; little accomplished in discussions.
20. Specific aspects—classroom laboratory, studio facilities; time of day of course.
21. Overall course—enjoyment of course; good course; course is beneficial; course opens up new ideas, stimulates my thinking.

Grading

22. Fairness—too easy, too strict; ambiguous test questions; changed grading procedures during semester; unclear grading procedures.

*L.A. Braskamp, J.C. Ory, and D.M. Pieper, "Student Written Comments: Dimensions of Instructional Quality," *Journal of Educational Psychology*, vol. 73, 1981, pp. 65-70.

perceive as being important to good teaching, or as characteristics descriptive of the "ideal" teacher. Feldman (1976), in a very thorough analysis of this line of research, identified 19 major characteristics based upon studies that included student nonstructured responses to open-ended questions and structured responses to the fixed-alternative items. (See Table 4.) These characteristics also parallel the major factors identified in the factor-analytic studies and the categories based on the content analysis of the student comments.

Feldman also determined the relationship between student overall evaluation of the instructor and student evaluation of the instructors for each of the dimensions or categories with the 10 highest rankings. The rankings of the characteristics based upon nonstructured responses and structured responses are definitely correlated, although not identically ranked. Feldman summarizes the results of these studies as follows:

> "Across studies using any of these types of measurement, stimulation of interest and clarity (understandableness) are the two dimensions that consistently highly associate with superior college teachers or teaching. Knowledge of subject matter is also highly characteristic of the superior teacher in effective teaching although somewhat less consistently so across studies. The other two consistently important characteristics are those that have been classified as instructor's preparation for (and organization of) the class and instructor's enthusiasm for the subject matter or for teaching." (p. 264)

Feldman further specifies that characteristics relating to the "regulative activities" of the instructor—e.g., class management, impartiality of evaluation, clarity of course, objectives and requirements, and difficulty of the course—are rated consistently lower in importance for college teaching. However, friendliness, concern and respect for students, helpfulness, and openness to others' opinions were among the most frequently mentioned characteristics when students freely described their best teachers or identified characteristics important to good teaching. However, when students responded to a preset list of characteristics, these three dimensions were less important. In another study, Goldsmid, Gruber, and Wilson (1977) found that superior teaching, as defined by those winning instructional awards on a large campus, were both competent and conscientious. They characterized the superior teacher as one who has "the hardness of the head and the softness of the heart."

Table 4

Characteristics of 'Ideal' and 'Best' College Teachers as Seen by College Students*

1. *Stimulation of interest:* "the instructor puts material across in an interesting way;" "the instructor gets students interested in the subject;" "it was easy to remain attentive;" "the teacher stimulated intellectual curiosity;" etc.

2. *Instructor's enthusiasm (for subject or for teaching):* "the instructor shows interest and enthusiasm in the subject;" "the instructor seems to enjoy teaching;" etc.

3. *Instructor's knowledge of subject matter:* "the instructor has a good command of the subject material;" "the teacher has a thorough knowledge, basic and current, of the subject;" "the instructor has good knowledge about or beyond the text book;" etc.

4. *Instructor's intellectual expansiveness (and intelligence):* "the instructor is well informed in all related fields;" "the instructor has respect for other subject areas and indicates their relationship to his or her own subject of presentation;" "the instructor exhibited a high degree of cultural attainment;" etc.

5. *Instructor's preparation; organization of the course:* "the instructor was well prepared for each day's lecture;" "the presentation of the material is well organized;" "the overall development of the course had good continuity;" etc.

6. *Clarity and understandableness:* "the instructor made clear explanations;" "the instructor interprets abstract ideas and theories clearly;" "the instructor makes good use of examples and illustrations to get across difficult points;" "the teacher effectively synthesizes and summarizes the material;" etc.

7. *Instructor's elocutionary skills:* "the instructor has a good vocal delivery;" "the teacher speaks distinctly, fluently, and without hesitation;" "the teacher varied the speech and tone of his or her voice;" etc.

8. *Instructor's sensitivity to, and concern with, class level and progress:* "the instructor was skilled in observing student reactions;" "the teacher was aware when students failed to keep up in class;" "the instructor teaches near the class level;" "the teacher takes an active personal interest in the progress of the class and shows a desire for students to learn;" etc.

9. *Clarity of course objectives and requirements:* "the purposes and policies of the course were made clear to the student;" "the instructor gave a clear idea of the student requirements;" "the teacher clearly defined student responsibilities in the course;" etc.

10. *Nature and value of the course material (including its usefulness and relevance):* "the instructor has the ability to apply material to real life;" "the instructor makes the course practical;" "there is worthwhile and informative material in lectures that doesn't duplicate the text;" "the course has excellent content;" etc.

11. *Nature and usefulness of supplementary materials and teaching aids:* "the homework assignments and supplementary readings were helpful in understanding the course;" "the instructor made good use of teaching aids such as films and other audiovisual materials;" etc.

12. *Difficulty of the course (and workload):* "the instructor covered the right amount of material;" "the difficulty of the course was reasonable;" "the pace of the course was not too fast or too slow;" "the amount of outside readings seemed appropriate;" etc.

13. *Instructor's fairness; impartiality of evaluation of students' quality of examinations:* "grading in the course was fair;" "the instructor has definite standards and is impartial in grading;" "the exams reflect material emphasized in the course;" "test questions were clear;" "coverage of subject matter on exams was comprehensive;" etc.

(continued)

14. *Classroom management:* "the instructor controls class discussion to prevent rambling and confusion;" "the instructor maintained a classroom atmosphere conducive to learning;" "students are allowed to participate in deciding the course content;" "the teacher did not 'rule with an iron hand';" etc.

15. *Nature, quality, and frequency of feedback to students:* "the instructor gave satisfactory feedback on graded material;" "criticism of papers was helpful to students;" "the instructor told students when they had done a good job;" "the teacher is prompt in returning tests and assignments;" etc.

16. *Encouragement of questions and discussion; instructor's openness to opinions of others:* "students felt free to ask questions or express opinions;" "the instructor stimulated class discussions;" "the teacher encouraged students to express differences of opinions and to evaluate each other's ideas;" "the instructor invited criticisms of his or her own ideas;" "the teacher appeared receptive to new ideas and the viewpoints of others;" etc.

17. *Intellectual challenge; encouragement of independent thought:* "this course challenged students intellectually;" "the teacher encouraged students to think out answers and follow up ideas;" "the teacher attempts to stimulate creativity;" etc.

18. *Instructor's concern and respect for students; friendliness of instructor:* "the instructor seems to have a genuine interest in and concern for students;" "the teacher took students seriously;" "the instructor established good rapport with students;" "the teacher was friendly toward all students;" etc.

19. *Instructor's availability and helpfulness:* "the instructor was willing to help students having difficulty;" "the instructor is willing to give individual attention;" "the teacher was available for consultation;" "the teacher was accessible to students outside of class;" etc.

*This listing is taken from K.A. Feldman's "The Superior College Teacher from the Student's View," *Research in Higher Education*, vol. 5, 1976, pp. 243-288. Used by permission of the author.

The last basis that has been used to examine effective instruction is to use colleague appraisals. The assessment of teaching by colleagues has included judgments on the course materials, tests, student achievement, and judgments based on classroom visits. The lists of components to judge are most often based on expert judgment with little empirical support. The lists have included high-inference dimensions, involving considerable judgment in addition to description. Very few systematic observation techniques—e.g., the one developed by Flanders—have been used or advocated to assess instruction in higher education. Most often global judgments by peers are written for promotion purposes. Agreement by peers may be reduced when the observation is unstructured, but this type of evaluation is often preferred.

Measurement of Teaching Effectiveness: A Critique

Thus far we have merely described the major dimensions of instruction. We have not yet examined the psychometric characteristics of the scales or instruments used to provide behavior indicators of these dimensions, nor have we assessed the usefulness of these scales in an evaluative context. That is, to what extent is the evaluative information collected by these measurement techniques useful either for instructor improvement or for administrators and other faculty who make personnel decisions. Thus we have three major questions that we need to ask: How reliable are the ratings obtained through structured observation? How valid is the information—i.e., for what purpose can it be used? How useful is the information based on structured observation techniques for decisions involving course improvement or for personnel matters?

Reliability. Reliability in evaluating instruction can take on different meanings, but the extent of agreement by students within a given class in judging or describing a teacher or course, and the stability of ratings, are the two most common interpretations. The desired level of reliability is somewhat dependent upon the purpose for which the information will be used—with use for promotion requiring a higher degree of reliability than for course improvement. Based upon the generalizability notion by Gilmore, Kane, and Naccarato (1978), it is recommended that ratings be obtained from five or more courses in which there are at least 15 students in each course taught by the same instructor. If an instructor teaches fewer than 15 students, then more

than five ratings should be used. The stability of student responses is relatively good over a short period of time, but as expected an increase in the amount of time between ratings will lower reliability (Overall and Marsh, 1978).

Validity. There are many issues involved in determining the extent to which a given set of data is valid. One is the concern over *confounding*—i.e., what characteristics of the course and the instructor serve as potential biases in the ratings of instruction? To what extent do secondary characteristics—i.e., characteristics of the instructor, the students, or the course that are not directly related to student learning—influence the rater's assessment? To what extent do improper administration procedures influence the rating?

The research findings are too complex and numerous to provide a thorough analysis of the studies in this paper, but I will briefly describe the generalizations which can serve as "red flags" in using ratings for course improvement or, more especially, for personnel decisions.

Student characteristics such as sex, academic major, academic ability, and age do not greatly influence ratings; however, expected grade point average may affect students' assessment of the instructor. In studies at the University of Illinois we have found that there is a correlation ($r = .42$) between students' rating of the instructor and expected grade (Brandenburg, Slinde, Batista, 1977). However, the interpretation of this correlation is not straightforward. Students who learn more also rate their teachers higher. On the other hand, teachers who are easy graders tend to get higher ratings.

Based on studies of instructor characteristics, academic rank does make a difference with teaching assistants as a group receiving lower ratings. Sex of the instructor, the teaching load, and research productivity are not related to effective instruction, although the relationships are often complex. Hoyt (1970) found no relationship for professors in the social sciences, whereas for natural science professors, a positive relationship between research productivity and student ratings was obtained. Instructor characteristics such as warmth, enthusiasm, and concern for students have been correlated with overall teaching ability. Centra (1980) found that years of teaching also made a difference on the ratings that teachers receive. First-year teachers generally receive the lowest ratings, and teachers with two to 12 years receive the highest ratings, after which the ratings begin to decline.

Course characteristics such as class size are important to con-

sider in interpreting student ratings. Small classes—those with less than 10 or 15 students—are the most highly rated. The relationships between class size and ratings is not always linear because there may be an optimal class size for discussion and creation of climate. A selection bias of asking the "best" teachers to teach large sections also complicates the interpretation of ratings. Subject matter can make some difference, since teachers in the field of humanities on the average receive higher ratings than those in social sciences and natural sciences. Also, students give higher ratings to elective courses or ones in their major area.

A very important issue about the validity of student ratings centers around the relationships between ratings and measures of student learning. That is, to what extent is there convergent validity between student ratings of instruction and the amount students learn on a classroom exam or some objective measure of course achievement? Cohen (1980) identified 41 independent validity studies in which multisection courses were used in the research design. The relationship between student achievement and overall instructor or course rating and the relationship between student achievement and specific dimensions (such as rapport and pedagogical skills) were both investigated. The average correlation between overall instructor rating and student achievement was .43, whereas the correlation between instructor and course rating and student achievement was .47. Lower correlations were found for relationships between learning and teacher skill and course structures. Correlations between student learning and student rapport, student/teacher interaction, feedback, and evaluation were only modest; no relationships between student ratings of course difficulty and student learning were obtained. This meta-analysis supports the conclusion reached by Frey (1978) that ratings of pedagogical skills such as planning, presentation, and clarity are less susceptible to bias than items that deal with the student/teacher relationships and instructor warmth.

Finally, this meta-analysis is also consistent with the recent advocacy by most experts in the field of evaluation of instruction at the postsecondary level. It is more defensible to use overall ratings of instructor or course such as "rate the course" or "rate the instructor" as evaluative information for use in personnel decisions than it is to use information that relates to the warmth of the faculty member and the type of relationships that students form with faculty members.

A third issue involves the convergent validity of ratings by

various groups of raters or data collected by different measurement techniques. The only two studies we could locate about convergent validity were two done by our office. We correlated students' objective ratings with written comments about the instructor and about the course in general (Braskamp, Ory, and Pieper, 1981). Students' ratings of the instructor are more highly correlated than student ratings of the course, given these two methods of data collection. The convergent validity between student ratings of overall effectiveness of the instructor based on those two methods was high ($r = .75$). In a second study (Ory, Braskamp, and Pieper, 1980), we intercorrelated ratings based on group interviews with students, student responses to fixed alternative items, and student written comments. Again the agreement of overall teaching ability was very high (correlations were over .90).

Recently a number of studies have been conducted on the convergent and discriminant validity of ratings provided by students, peers, alumni, and instructors themselves. While there have been many who have been suspect on the usefulness and validity of self-ratings, Braskamp, Caulley, and Costin (1979) found that self-ratings demonstrate good convergent validity on student involvement, teacher support, and teacher skill for instructors in their second semester of teaching. Self-ratings were also not unusually high or more favorable than student ratings. Alumni also tend to rate teachers similarly (Centra, 1974; Overall and Marsh, 1978).

There is also some convergent validity between colleague ratings and student ratings. When colleagues are asked to rate their peers without any classroom visit, they tend to agree with students on the general overall teaching ability of their colleagues (Blackburn and Clark, 1975; Doyle and Crichton, 1978). However, the colleagues may be basing their judgment on what students tell them. In one study (Centra, 1975), in which peer ratings were based on two classroom visits of faculty within and without the department, colleagues did not agree with each other and their ratings were not highly related to student ratings when overall ability was being evaluated. However, colleagues and students reasonably agreed on items relating to specific instructional practices. Centra also concluded that colleague judgments of instructor skill and general teaching effectiveness based upon two classroom visits was not sufficiently reliable to warrant their use for promotion purposes.

Conclusion

The research on instruction at the collegiate level has centered on personality characteristics of the instructor and course structure. By comparison the student has not been studied except to use student achievement to check on the validity of student ratings of instructor characteristics such as classroom presentation skills, warmth, and course demands. However, the research suggests that effective teaching requires an instructor to be competent and to be concerned and enthusiastic about student learning. Competence includes a command of the subject matter and the ability to organize the course to enhance student learning. Concern includes the ability of the instructor to provide a supportive, encouraging, and stimulating learning environment and to present the content in an understandable manner from the student's perspective. These components are not very behaviorist or concrete, although specific teacher competencies can be identified as exemplars. Nevertheless, the research findings based on collegiate instruction are not definitive enough to ignore the principles stated earlier in this paper—that teaching is too dynamic to be able to capture it by one measurement technique.

References

Aubrecht, J.D. "Are Student Ratings of Teacher Effectiveness Valid" (ICEA Paper No. 2). Manhattan, Kan.: Kansas State University, Center for Faculty Evaluation and Development, 1979.

Brandenburg, D.C., J.A. Slinde, and E.E. Batista. "Student Ratings of Instruction: Validity and Normative Interpretations," *Journal of Research in Higher Education*, vol. 7, 1977, pp. 67-78.

Brandenburg, D.C., S. Derry, and D.D. Hengstler. "Validation of an Item Classification Scheme for a Student Rating Item Catalog," a paper presented at the National Council on Measurement in Education annual meeting, Toronto, April 1978.

Braskamp, L.A. and R.D. Brown. *Utilization of Evaluative Information*. San Francisco: Jossey-Bass, 1980.

Braskamp, L.A., J.C. Ory, and D.M. Pieper. "Student Written Comments: Dimensions of Instructional Quality," *Journal of Educational Psychology*, vol. 73, 1981, pp. 65-70.

Braskamp, L.A., D.N. Caulley, and F. Costin. "Student Ratings and Instructor Self-Ratings and their Relationship to Student Achievement," *American Educational Research Journal*, vol. 16, 1979, pp. 295-306.

Centra, J.A. *How Universities Evaluate Faculty Performance: A Survey of Department Heads*. GREB Publication No. 75-5. Princeton, N.J.: Educational Testing Service, 1977.

_____. "The Relationship Between Student and Alumni Ratings of Teachers," *Educational and Psychological Measurement*, vol. 34, 1974, pp. 321-326.

_____. "Colleagues as Raters of Classroom Instruction," *Journal of Higher Education*, vol. 46, 1975, pp. 327-337.

_____. *Determining Faculty Performance*. San Francisco: Jossey-Bass, 1980.

Cohen, P.A. "A Meta-Analysis of the Relationship Between Student Ratings of Instruction and Student Achievement." Ph.D. dissertation. Ann Arbor, Mich.: University of Michigan, 1980.

_____. "Effectiveness of Student Rating Feedback for Improving College Instruction: A Meta-Analysis of Findings," *Research in Higher Education*, vol. 13, 1980, pp. 321-341.

Costin, F., W.T. Greenough, and R.J. Menges. "Student Ratings of College Teaching: Reliability, Validity, and Usefulness," *Review of Educational Research*, vol. 41, 1971, pp. 511-535.

Diamond, N., G. Sharp, and J.C. Ory. *Improving Your Lecturing*. Urbana, Ill.: Office of Instructional Resources, University of Illinois, 1978.

Doyle, K.O., Jr. *Student Evaluation of Instruction*. Lexington, Mass.: Lexington Books, D.C. Heath, 1975.

Doyle, K.O., Jr. and L.A. Crichton. "Student, Peer, and Self-Evaluation of College Instruction," *Journal of Educational Psychology*, vol. 70, 1978, pp. 815-826.

Erickson, G.R. and B.L. Erickson. "Improving College Teaching: An Evaluation of a Teaching Consultation Procedure," *Journal of Higher Education*, vol. 50, 1979, pp. 670-683.

Feldman, K.A. "The Superior College Teacher from the Students' View," *Research in Higher Education*, vol. 5, 1976, pp. 243-288.

Frey, P.W. "A Two-Dimensional Analysis of Student Ratings of Instruction," *Research in Higher Education*, vol. 9, 1978, pp. 69-91.

Gillmore, G.M., M.T. Kane, and R.W. Naccarato. "The Generalizability of Student Ratings of Instruction: Estimation of the Teacher and Course Components," *Journal of Educational Measurement*, vol. 15, 1978, pp. 1-13.

Glass, G.V. "A Paradox about Excellence of Schools and the People in Them," *Educational Researcher*, vol. 4, 1975, pp. 9-13.

Goldsmid, C.A., M.E. Gruber, and E.K. Wilson. "Perceived Attributes of Superior Teachers (PAST): An Inquiry into the Giving of Teacher Awards," *American Educational Research Journal*, vol. 14, 1977, pp. 423-440.

Hildebrand, M., R.C. Wilson, and E.R. Dienst. *Evaluating University Teaching*. Berkeley, Calif.: Center for Research and Development in Higher Education, University of California, 1971.

Hoyt, D.P. *Instructional Effectiveness—Interrelationship with Publication Record and Monetary Reward*. Manhattan, Kan.: Kansas State University, Office of Educational Research, 1970.

Hoyt, D.P. and G.S. Howard. "The Evaluation of Faculty Development Programs," *Research in Higher Education*, vol. 8, 1978, pp. 25-38.

Kulik, J.A. and C.C. Kulik. "College Teaching," in P.L. Peterson and H.J. Walberg, eds., *Research on Teaching: Concepts, Findings, and Implications*. Berkeley, Calif.: McCutchan, 1980.

Kulik, J.A. and W.J. McKeachie. "The Evaluation of Teachers in Higher Education," in F.N. Kerlinger, ed., *Review of Research in Education*. Itasca, Ill.: F.E. Peacock, 1975.

Marsh, H.W., J.U. Overall, and S.P. Kesler. "Validity of Student Evaluations of Instructional Effectiveness: A Comparison of Faculty Self-Evaluations and Evaluations by their Students," *Journal of Educational Psychology*, vol. 71, 1979, pp. 149-160.

McKeachie, W.J. "Psychology in America's Bicentennial Year," *American Psychologist*, vol. 31, 1976, pp. 819-833.

_____. "Student Ratings of Faculty: A Reprise," *Academe*, vol. 65, 1979, pp. 384-397.

Instructor and Course Evaluation System (ICES) Newsletter, No. 1-4. Urbana, Ill.: Office of Instructional Resources, University of Illinois, 1978.

Ory, J.C., L.A. Braskamp, and D.M. Pieper. "The Congruency of Student Evaluative Information Collected by Three Methods," *Journal of Educational Psychology*, vol. 72, 1980, pp. 181-185.

Ory, J.C. and L.A. Braskamp. "Faculty Perceptions of the Quality and Usefulness of Three Types of Evaluative Information," *Research in Higher Education*, in press.

Overall, J.U. and H.W. Marsh. "Long-Term Stability of Students' Evaluation of Instruction: A Longitudinal Study." A paper presented at the Association for Institutional Research meeting, Houston, 1978.

Peterson, P.L. and H.J. Walberg, eds., *Research on Teaching: Concepts, Findings, and Implications*. Berkeley, Calif.: McCutchan, 1980.

Rosenshine, B. "Academic Engaged Time, Content Covered, and Direct Instruction," *Journal of Education*, vol. 160, 1978, pp. 38-66.

Seldin, P. *Successful Faculty Evaluation Programs*. Crugers, N.Y.: Coventry Press, 1980.

Components of Teaching Based on Values and Valuing

Larry A. Braskamp

The purpose of this presentation is to examine the role and contribution of values and valuing in the assessment of instruction. Once we use the words *effective instruction,* as we have so far in our assessment of instruction, we are implying a degree of goodness, quality, worth, and value to instruction. Effective instruction involves judgment and thus is more than description.

An analysis of valuing is very amorphous, almost without boundary. To begin our thinking about the contribution of valuing to the assessment of instruction, I shall first define *value.* According to Webster, three definitions of the verb are: 1) to estimate the value of; set a price for; determine the worth of; appraise; 2) to place a certain estimate or worth on in a scale of values: as, I value health above wealth; 3) to think highly of; esteem; prize; as, I value your friendship. In a recent presentation Bill Gephart (1980) gave several definitions of value used both as a noun and a verb. One definition he gave was "caring for." As such, it represents an internal affective state. At times it may connote action, action viewed as the result of "caring for." Values can be thought of as internal predispositions or preferences. They are primarily affective, but they represent more than an attitude toward an object. They can serve as guiding principles of what we consider good for ourselves and for the society in which we live.

The valuing process encompasses choices and actors. We are the actors, and as human beings we have the intellectual, social, and emotional capacity and freedom to choose. Valuing is fundamentally a human activity. To be human is to have the potential to be free to make choices. A robot or animal does not have this freedom of choice. Our freedom is, of course, limited and constrained by our own capacity and the rules, norms, and expectations of the society in which we live. We all like to view ourselves as having a will, whether we think it is entirely free, partially determined by societal forces, or predetermined to some extent by some supreme authority.

The human acts are choices, choices among alternative values. Our choices are based on what we consider to be best for us and our society.

Our personal values are highly individualistic and are difficult to specify. Often they are implicit but nevertheless powerful. They also change as we grow or develop, change as much in complexity and integration as well as in kind or type.

Our personal values serve as the basis for making judgments of worth about particular programs, such as those in our educational system. For example, our judgment about a particular school's successes in teaching the basic skills of reading may well reflect the degree of importance we give reading skills—i.e., the value we give to reading in the overall curriculum.

We have to make many choices—political, social, ethical, economic, educational, and religious. Thus our values are of many types, and we may not always be able to integrate them very well. Some may even be contradictory.

In making value choices, we often employ a means-end analysis. We have notions of what our goals and ends are, and we select means to try to maximize those goals. For example, we force children to attend school since we value the end result—educated people. However, sometimes the means and goals cannot be distinguished very clearly, and means often serve as intermediate goals. Other times the means can serve as ends, or goals. They have intrinsic value, such as requiring students to attend an art show or participate in competitive athletics. We often take for granted our objectives and goals and concentrate on the means to maximize the goals.

To this point I have described the valuing process as a sequence involving three distinct types of choices we make as humans and as professionals. They are: 1) values as preferences, or principles; 2) judgments of worth or merit based on a set of criteria; and 3) decisions to act upon the judgments. At each stage of this sequence choices are made, i.e., selections from among alternatives are necessary. Each stage represents a different level of abstraction and comprehensiveness, ranging from the general to the specific and from the philosophical to the concrete.

Up until now the discussion has focused on our own personal set of values and how we may think of them and how we may defend them. But evaluation of instruction that takes place within a public or private institution is no longer a private matter. Evaluation is a social and public activity. The evaluation of a program, instructor, course, or curriculum involves making a judgment of worth, hopefully based on some collective value system.

Our personal choices may not always be the same as the collective one. For example, as a school administrator I may value

music education very much, and I may judge the program at a local school to be very inadequate based on the evidence collected about the music program. I may also desire to have a symphony band at the elementary school level. However, as a servant of the community, my personal choices may not be implemented or even seriously considered. The athletic program may be the choice of the citizens. At each stage there are many possible value positions, judgments of worth, and administrative decisions. The formal evaluation of a specific program usually focuses on the second stage, often without much regard to the underlying collective value system of the affected persons, and without much regard to the economic and political realities underlying the adoption of a formal evaluation.

Given our decision to evaluate (which is also a value choice), what are some of the issues we should consider in the evaluation of instruction? There are four major issues that I want to discuss briefly. They are 1) role of education in society, 2) professional preferences, 3) the distinction between worth and merit, and 4) demonstration of accountability.

The Role of Education in Society

The evaluation of instruction is not accomplished in isolation; its context is vital. Schools perform a service to society and its citizens. Our societal norms, job entry skills, and cultural or religious heritage are learned by those attending an educational system.

If we also take seriously our pluralist and democratic society, our assessment of education becomes even more difficult. House (1980) writes: "A public evaluation must be normatively correct.... Public evaluation should be democratic, fair, and ultimately based on the moral values of equality, autonomy, impartiality, and reciprocity" (p. 256). He further states that in many evaluations, the focus is too narrow: "Obligations to the larger society are frequently overlooked. The evaluation of public programs is clearly a situation in which such obligations do exist, and the concept of validity must be extended to apply to such a significant social practice" (p. 257).

An evaluation should "recognize a societal and public interest beyond the private interests of individuals" (p. 141). An evaluation is contingent upon the reference group and the values of the group. "The proper reference group for an evaluation...are all those who are affected by the program or policy" (p. 142). For

example, suppose we wish to evaluate the quality of instruction at a community college. This evaluation could be very direct and relatively straightforward if we accept the stated instructional goals for a particular course and design an evaluation to measure the extent to which students acquire the desired ends, i.e., select some form of goal-based evaluation strategy. But this type of evaluation is a rather mechanical process, and to many evaluators it misses the essence of evaluation—namely, assignment of values.

Michael Scriven, in his early writing on evaluation, suggested that the goals themselves should also be evaluated and more recently has advocated that goal-free evaluations should be conducted. This means that a consumer-based evaluation using consumer needs as the standard for assessing value should be employed. If we follow this strategy we are moving further away from the classroom and into society to find our criteria and standards to use in evaluation. Evaluation in this context raises new problems, but they are problems that should be addressed in a thorough evaluation.

Michael Apple (1974) also emphasizes the importance of examining educational evaluation from a broader perspective than just from the students:

> I should begin by stating that certain types of performance, certain forms of knowledge, certain dispositions, achievements, and propensities are not necessarily good in and of themselves. Rather, they are made so because of specific taken-for-granted assumptions. Thus, their values are relative and temporal conditions. In order to make this clear, another significant but often unrecognized fact should be mentioned here. The guiding principles of evaluation—conceptions of achievement, of success or failure, and so on—are social constructs. They are not inherent in individuals or groups of people. Instead, they are instances of the application of identifiable social rules about what is to be considered good or bad performance. . . . Educators must examine the ideological and political uses of evaluation and the place of the school in a larger social setting if they are to uncover what evaluation is actually about (pp. 12-13).

Professional Preferences

Evaluators plan and conduct evaluations that often reflect their values on the validity, reliability, and credibility of evaluative information. They may select certain types of evaluation strategies and types of data to collect since they consider

these designs and information to best fulfill their notions of what is objective, reliable, and credible information. For example, the use of scales and objective measurement procedures often reflect credence they give to rather formal operations and representations of reality. This type of empirical evidence is given a certain type of trustworthiness by its standardization and objectivity. If we contrast this type of information with information collected in a naturalist form of evaluation—one in which human judgment and interpretation are based solely on the insight, intuition, and reasoning of the evaluator—we have two quite differing types of evaluative information. If we accept achievement test scores as the empirical evidence for our evaluation, we are implying that these scores represent social and educational worth. If we accept judgments of worth from multiple audiences, we are defining worth differently.

Our professional preferences also influence the way we interpret the available data (Krathwohl, 1980). We don't integrate, summarize, or draw conclusions from an empty mind set. Our conceptual framework colors the type of information we seek, and it helps to determine our willingness to accept data as confirming or refuting our hypotheses.

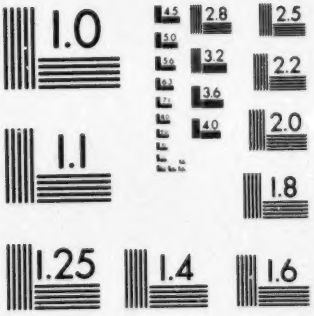
The choice of the evaluation strategy also reflects our values regarding the importance of those affected by the evaluation. I tend to like Cronbach's (1980) position that all those affected by the program or policy should be included in the evaluation. Thus the primary task of the evaluators is to uncover the value positions of the various audiences, especially those with unpopular, minority opinions.

Different evaluations have different biases toward the weight given to the various criteria used in an evaluation and the type of empirical evidence that best serves as a proxie or indicator of the criteria. Some evaluators wish to be very reductionist, by selecting only a few quantifiable variables to serve as the empirical basis for evaluation, whereas others prefer to use more subjective information reflecting implicit criteria.

Worth and Merit

Recently a distinction has been drawn between worth and merit, two facets of value. Based on an earlier paper by Scriven (1978), Lincoln and Guba (1980) distinguish worth and merit by the extent to which the value of something is influenced by its context. Merit is essentially context free, but worth can only be

PG



MICROCOPY RESOLUTION TEST CHART

determined in relation to context. The assessment of worth requires knowledge about "local cultural, political, and value factors" (p. 61). This distinction is important in evaluating faculty. The determination of merit of a faculty would most often include the person's teaching performance within the classroom. An evaluation including these considerations is relatively context free. Historically, those in higher education have made personnel decisions based on merit. In judging the *worth* of a faculty member, however, we need to consider the contribution made to a local institution such as teaching a high demand area, service work, collegiality, etc. Two professors may have the same merit ratings according to a selected set of criteria, but their worth may vary considerably depending on the local needs of institutions. Furthermore, worth—being context dependent—may vary depending upon the institutional needs and resources at a given time. One scholar in a department may be all an institution can afford, and thus the worth of any faculty member is contingent upon the current mix of faculty in the department.

Merit can be determined by using absolute and comparative standards. Since merit primarily represents intrinsic value, it can be judged by the extent to which the person being evaluated meets a certain set of standards, e.g., a group of peers evaluating one's ability to speak. On the other hand, peers may rank a group of professors on their elocution skills in the classroom, which requires *comparative* evaluation. Worth is largely determined by comparing the person being evaluated with others who are similarly affected by the evaluation. In this case the selection of the appropriate judgment or reference group is the key factor. In evaluating one's teaching performance several reference groups can be employed, e.g., students, peers within the department, peers within the institution, governing board, legislatures, and citizens. The selection of the reference group will determine the criteria for assessing merit and may likely influence the judgment of merit of the person being evaluated.

Accountability

Since education is a public enterprise, those in education are accountable for its success or failure. That is, educators are to be responsible. This responsibility is focused by the following three questions: To whom is one responsible? For what is one responsible? By what means will accountability be demonstrated?

Evaluation is often a means of demonstrating accountability.

Evaluations—as assessments of worth and merit—are used to communicate to designated publics the quality of programs or policies and the effectiveness of personnel. But, an overemphasis on strictly data-based evaluations can have significant negative side effects. In an earlier paper, Mel George and I (1978) discussed one problem with evaluation in higher education:

> Of even more fundamental significance, however, is that accountability mechanisms can change the fundamental character of the institution being held accountable, and often such a wide-ranging effect is not even recognized, much less intended. It's a little like the Heisenberg uncertainty principle, which states that because of the dual wave and particle nature of matter and energy, one cannot measure with arbitrary precision certain physical quantities without modifying the behavior of the system being observed. For instance, to measure accurately the position of an electron requires energy (say, light in a microscope) that modifies the momentum of the electron to such an extent that the momentum cannot be known with arbitrary accuracy. And the greater the precision demanded for the measurement of the position, the more uncertain becomes the momentum.

The Role of Values in the Evaluation of Instruction

I have argued that one's personal and professional values—in conjunction with a set of collective values—influence an evaluation. The questions that follow—questions that I think anyone ought to ask in preparation for evaluating instructors—reflect the relationship between values and evaluation.

1. *What are your own personal values?* What do you care about? What are the bases for your values, i.e., to what do you attribute them; how do you justify them?
2. *To what reference groups is the evaluation directed?* As a public servant, what are your responsibilities to demonstrate accountability in terms of relevant audiences or interest groups?
3. *Who decides what is to be evaluated?* What publics are to be included in the evaluation process? Whose values will be listened to?

   Evaluation theorists differ on the degree to which evaluators should actively engage audiences in the

evaluating process. Patton (1978) and Stake (1975) advocate heavy involvement to maximize utilization and meet information needs of the audiences. On the other side, Scriven proposes a more independent role for the evaluator, independent of the program administrator and a general consumer spokesperson. The evaluator has more autonomy and authority; thus his or her value system can play a significant role in the evaluation process. Stake also desires the evaluator to be influential, but the influence is through interaction by making the decision makers more aware of their own values, aware of the program's worth from the perspectives of other affected audiences. Krathwohl (1980) states that there are many positions in between, positions that nevertheless allow the evaluator to be proactive.

4. *What evaluation strategy are you employing in the formal evaluation?* Are the methods and data you collect reflecting the importance you place on the means and ends of the programs being evaluated? How explicit do the evaluation criteria have to be? For example, if your evaluation includes only measures of student achievement, then the audience has to conclude that student achievement is what is valued.
5. *What evaluative information or evidence is to be collected?* How well does the information represent that which is being evaluated? The choice of the measurement technique reflects the importance given to various types of information. For example, does a 20-item scale on self-esteem adequately represent the students' social growth after they have gone through a workshop on self-development? If you use structured observation techniques for classroom observation, you are capturing a certain segment of the classroom activity. Some observation techniques (e.g., Flanders) emphasize verbal behavior in the classroom, noting, for example, who speaks most often, or noting the predominant mode of verbal presenting or interaction (whether presenting or questioning). Nonverbal behaviors of teachers and students, quality of presentation such as integration and clarity of presentation, and amount of student learning are not assessed.
6. *How much weight is to be given to each piece of evaluative information?* If an overall index of effectiveness is needed, what procedures for integration and summarization are followed?

7. *To what end is the evaluation to be used?* Is the purpose of the evaluation to help those in authority to acquire a better understanding of the program, to allow them to reflect and gain new insights into the strengths and weaknesses of the program, or to provide a summary index of worth or merit for a specified decision?
8. *Finally what are your values about the merit or worth of an evaluation?* What price are you willing to pay for an evaluation? How committed are you to having evaluation be the means by which organizational change is brought about? How willing are you to have your evaluations set the standards for excellence?

Your answers to these questions should reflect your own values about evaluation.

References

Apple, M.W. "The Process and Ideology of Valuing in Educational Settings," in M.W. Apple, M.J. Subkoviak, and H.S. Luflin, eds., *Educational Evaluation: Analysis and Responsibility*. Berkeley, Calif.: McCutchan, 1974.

Cronbach, L.J. *Design of Evaluations*. Stanford, Calif.: Stanford Evaluation Consortium, 1980.

George, M.D. and L.A. Braskamp. "Universities, Accountability, and the Uncertainty Principle," *Educational Record*, vol. 59, 1978, pp. 345-366.

Gephart, W.J. "Values, Valuing, and Evaluation," paper presented at the annual meeting of the Evaluation Network, Memphis, Tenn., September 1980.

Gowin, D.B. and J. Millman. "Meta-Evaluation and a Direction for Research on Evaluation," *CEDR Quarterly*, vol. 11, no. 4, 1978, pp. 3-6.

House, E.R. *Evaluating with Validity*. Beverly Hills, Calif.: Sage Publications, 1980.

Krathwohl, D.R. "The Myth of Value-Free Evaluation," *Educational Evaluation and Policy Analysis*, vol. 2, 1980, pp. 37-46.

Lincoln, Y.S. and E.G. Guba. "The Distinction Between Merit and Worth in Evaluation," *Educational Evaluation and Policy Analysis*, vol. 2, 1980, pp. 61-72.

Patton, M.Q. *Utilization-Focused Evaluation*. Beverly Hills, Calif.: Sage Publications, 1978.

Scriven, M. *Evaluation Bias and Its Control*. Kalamazoo, Mich.: Western Michigan University, College of Education, Evaluation Center, 1975.

_____. "Merit vs. Value," *Evaluation News*, no. 8, December 1978, pp. 20-29.

Stake, R., ed., *Evaluating the Arts in Education: A Responsive Approach*. Columbus, Ohio: Charles E. Merrill, 1975.

Guidelines for Using Rating Scales and Checklists

Larry A. Braskamp

In the preceding two papers I emphasized the importance of using multiple resources and varied methods in teacher evaluations. To carry out a comprehensive evaluation of faculty members and their courses, one should collect evaluative information from a variety of publics, such as students and colleagues, and more than one method or instrument should be employed—e.g., objective rating scales, interviews, videotapes, classroom visits, content analysis, syllabuses, and examinations.

This approach is advocated because our research on teaching is inconclusive, and our values do influence our evaluation of instruction. Images and perceptions of effective instruction vary. We ought to encourage a variety of teaching methods, and our evaluations ought to reinforce that diversity, not inhibit it. The convergent validity of evaluative information collected from different sources via different methods is encouraging, but there is not a one-to-one correspondence, given the variety of sources and the evaluation instrument. To be fair, one needs to include more than one piece of evidence. Within this context I will list some guidelines that are important in the evaluation of faculty when the evidence is to be gathered from objective rating scales—whether from students, colleagues, or the instructors themselves.

1. *There is an inherent paradox in teacher evaluation.* Gene Glass (1975) has aptly described this paradox. If we are interested in the pursuit of excellence, it can be achieved if certain conditions are met. First, teachers must feel that their failures are understood and that their work is not being continuously scrutinized by others. Conversely, individuals move toward excellence by specifying goals, and feedback is provided about their progress toward achieving the goals through evaluation. Both positions seem valid, but they are antithetical. Given the history of faculty autonomy and independence, and the need for experimentation in higher education, we face a dilemma about the extent to which evaluation, particularly that which is externally imposed, should be practiced. As employees of institutions that receive funds from the general public, regardless of whether the institutions are private or public, faculty have an obligation

to demonstrate their accountability. Through evaluation, both the individual instructor and the institution can benefit. On the other hand, faculty members are professionals. As such, they need freedom to experiment and evaluate themselves. They also need to be able to explore, create, and teach their classes in a personalized way in order to achieve diversity within higher education. Thus how much can the public respect the private world of the instructor's classroom and trust professionals to monitor themselves, and how much can they require faculty to demonstrate accountability? There is another twist to this paradox that goes like this: External evaluation of one's work should lead toward something, and that something is usually a form of external reinforcement or reward--e.g., a salary increase. Those who develop performance appraisal systems in industry emphasize the connection between performance, evaluation, and reward for the level of performance. If the linkage among these three processes is not held up, the changes of reward influencing increased performance is reduced (Steers, 1977). However, there is also some evidence from experimental studies that an introduction of external assessment and reward for performance may reduce the intrinsic satisfaction of actually doing it. One can have his or her motivation to perform an inherently pleasing task reduced by the introduction of an external evaluation and reward structure. The critical issue is not whether evaluation should be undertaken but the manner in which a faculty member is to be evaluated. The potential *dysfunction* of evaluation needs to be examined as carefully as the function of evaluation. We want to avoid the error of doing too much evaluation, or as Hoyt (1980) says, committing the "overkill error."

2. *Evaluation serves many purposes and functions, and they must be clearly differentiated early in the evaluation process.* Evaluation of instruction serves three primary purposes: a) assists the individual instructor to improve as a teacher; b) provides information to colleagues and administrators for decisions about promotion, tenure, annual salary increases, and new responsibilities; and c) provides information to students for course selection. Improvement of instruction is the function most often talked about when evaluation is discussed. In day-to-day practice, however, it is clear that administrators are increasingly using evaluation as a basis for personnel decisions. Administrators have required more information in recent years because of pressures resulting from tighter budgets and increased faculty litigation over personnel issues.

The purpose of the evaluation makes a difference with respect to the type of information that is appropriate. Generally, there is a consensus in higher education that general overall ratings of instructor ability and instructor pedagogical skills are more appropriate for use in personnel decisions than are student assessments of student/faculty relationships. The rationale rests heavily on two factors. First, external influences—e.g., class size, faculty characteristics, and whether the course was required or was an elective—do not seem to confound the ratings of an instructor's pedagogical skills as much as ratings of relationships do. Second, general overall ratings are more highly correlated with student achievement, the criterion considered by many to be the one we should be using to assess teacher performance. If the purpose is improvement, highly detailed diagnostic information is more useful.

3. *Evaluative information must meet adequate psychometric and technical standards.* Technical quality refers to the extent to which the information is comprehensive, reliable, valid, and free from potential biases and irrelevant considerations. The level of technical quality can vary depending upon the purposes of evaluation. Reliability should be higher for personnel decisions than for course improvement decisions.

As a general rule, personnel decisions about individual faculty members should be based upon at least five courses in which at least 15 students were enrolled. If the number of students is larger, this rule can be relaxed. If one has taught a large number of small classes, adjustments in the ratings should be considered since ratings of small classes tend to be unusually high.

Standard operating practices (e.g., standardized administration procedures, parallel reporting formats, and consistent interpretation of the results) need to be designed and followed. If normative information is to be used, the appropriateness of the reference group or normative group to which one is being compared is critical. Because of the differences in ratings due to factors other than teacher competence (e.g., class size, rank of instructor, major field of study, and required/elective nature of the class), adjustments need to be made either by using specified smaller reference groups or recognizing the appropriateness of the norms or reference group when the results are interpreted, especially for promotion purposes. Also use a range to highlight the unreliability of ratings.

4. *The evaluative information must be fair.* Fairness refers to the extent to which the evidence adequately represents the teaching/learning activities, the complexity of teaching, and a recognition of the possibility that different interpretations of

teaching may be the result of using different methods of evaluation. Fairness also refers to the extent to which the information adequately represents the criteria. This principle involves values because the selection of the criteria are as value based as they are grounded on research evidence. Perhaps a good way to look at whether information is fair is to ask ourselves four basic questions: 1) What is to be evaluated? Is the subject matter of the evaluation adequately covered by the items selected for the evaluation instrument? If not, the probability of having a valid evaluation is reduced proportionately. 2) What is the emphasis on each piece of evidence? What type of differential rating or importance is given to the various types of information? If the information is to be used diagnostically, the weighting process may not be as important as if it were used for personnel decision making. The establishment of *a priori* guidelines for weighting evidence is important in establishing ground rules. Potential users of the information then will at least know the rules of the game which should enhance fairness. 3) What measurement procedures are to be employed in requesting the desired information? The principle of adequate psychometric and technical quality is appropriate here. 4) How are errors, biases, and subjective judgments to be minimized or at least accounted for? Perhaps the best safeguard for this is the use of the multiple-method, multiple-source approach to evaluation. Another is to fully describe before you interpret or judge.

5. *The evaluation must have credibility to the intended users.* Credibility must generally be earned. Unlike validity and reliability, it is more than a technical matter. It involves politics; it involves mutual trust between those being evaluated and those using the evaluative information. The external constraints under which the evaluation is to be implemented become important here.

Credibility is assessed from the perspective of the potential users more so than it is in the perception of the evaluator. If the information is not viewed as credible, information that is technically sound will still not be as highly regarded. There are often tradeoffs between credibility and validity; that is, there is not a perfect relationship between valid and credible information. Furthermore, the same information will not be viewed as credible for all purposes. As our simulation showed (Ory and Braskamp, in press), qualitative and subjective information by students was viewed as more useful and credible for course improvement purposes, whereas the more traditional statistical data were viewed as more credible and believable for promotion purposes.

Hoyt (1980) describes three principles relating to credibility that are worth mentioning: The first one is labeled the "uniqueness principle." It involves the sensitivity of the evaluation toward differences among people and situations. Identical procedures and standards cannot be applied mechanically to both assistant and full professors, nor for professors at a variety of different institutions. The second is labeled the "power principle." This principle relates to the degree to which those affected have a voice in designing the evaluation system. If people feel powerless they will tend to resist and resent the process regardless of the technical quality of the evaluative information. To enhance credibility, Hoyt suggests involving those to be evaluated in determining the rules of the game. The third is the "reinforcement principle." If evaluation is seen primarily as something to be feared—i.e., "I stand to gain very little, but I could lose much"—then the credibility of the evaluation will not be high.

6. *The evaluation process should lead toward faculty self-development.* This guideline is based on the "reinforcement principle." The formative roles of evaluation should be stressed in any evaluation. The information must be helpful to the instructor or the user if it is to be functional. For example, just knowing that one is a poor teacher is definitely not very helpful.

McKeachie (1976), in his presidential address to the American Psychological Association, in which he reviewed the history of psychology in our first 200 years, gives three reasons why feedback of student ratings on instructor improvement has not been very encouraging. If information is to be useful, it must meet three conditions. First, the information has to be explicit—that is, information about specific strengths and weaknesses may be needed. Second, the user has to be sufficiently motivated to use this information and to act upon it. Third, the information must provide alternative strategies for improvement. If instructors are informed that they are doing a poor job, but do not know how to improve and are not given any alternative strategies for improvement, the end result is discouragement, resentment, and perhaps even a poorer performance. Thus an evaluation should be aimed toward some improvement if we expect growth and change to occur.

There are times, of course, when only summative information about promotion and tenure is desired. But in the entire scheme of evaluation within an institution, the tenure decision, though perhaps the most important one in the life of an individual faculty member, should be a minor part of the total evaluative efforts

of an individual faculty member. The implication of this principle is that faculty members should become better self-evaluators. It is too expensive and ineffective to have someone else in charge of evaluation.

7. *Reporting and communication channels are as important as the evaluative information collected.* This principle is an extension of the previous guideline. Data alone are not very useful. To enhance usefulness of data we need to be more concerned with the reporting and communication channels.

In a number of studies on the utilization of evaluative information, the key variable is the extent to which there is a personal relationship established between the evaluator and the user. Evaluation will be useful if communication is not a one-way process. Rather, communication should be an ongoing dialogue between the evaluator and the user. It should occur both informally, formally, and continuously. While it may not be necessary for an external evaluator to personally present the information to an individual faculty member on a continuous basis, the importance of the consultative arrangement in providing evaluative information back to the faculty should be stressed. Provisions for personal contact between the faculty and the evaluator (department head, colleague) and opportunities for a faculty member to react to the evaluation should be made. Data from scales and surveys presented in a statistical report are not very exciting and provocative. However, this information can serve as a catalyst for discussion. Thus for formative uses, dialogue and discussions should be a part of the feedback process.

If the purpose of the evaluation is for summative personnel decision making, the communication channels should be different from those that are provided for self-improvement. The channels should be spelled out in advance and they should be followed judiciously. Provisions for checks and balances of information should be built into the procedures. While it is impossible to present draft reports of confidential evaluation to the individual faculty member, committees responsible for evaluating faculty members should be extremely prudent in checking out their sources and looking for inconsistencies and discrepancies and to follow up on evaluations that seem particularly out of line.

8. *The information must be useful to the institution.* The evaluations need to be accessible to those involved in the decision-making process. In addition, the information must be understandable to the users. The reporting format can influence readability. For example, eliminating as many numbers as possible makes the report more readable.

The degree of usefulness of any evaluation has at least two frames of reference, individual and institutional. Each perspective may be different; in some cases they may conflict. For example, in evaluating the usefulness of colleague ratings, we need to consider seriously the extent to which mutual trust among departmental faculty is impaired or enhanced, the worthwhileness of the time required to implement colleague ratings, and changes that an instructor may make by working with a colleague.

9. *The evaluation process must fit into the governance and organizational structure of the institution.* The evaluation must complement, not usurp the existing organization and flow of decision making within an institution. Specific procedures for information flow, involvement of faculty in the evaluation process, and reconciling the needs of individual faculty members with the needs of the institution should be spelled out.

10. *Feasibility of the evaluation process must be considered.* A comprehensive evaluation process meeting most of the previous guidelines may not be possible due to lack of time or financial resources. Return on investment must also be considered when discussing alternative evaluation processes. Very often, the ideal will not be possible. Given the constraints of time, politics, and institutional needs, compromise may well have to be considered. In the absence of a willingness to compromise, the evaluation plan will never leave the paper on which it is written.

References

Glass, G.V. "A Paradox about Excellence of Schools and the People in Them," *Educational Researcher,* vol. 4, 1975, pp. 9-13.

Hoyt, D.P. "Principles of Evaluation," a paper delivered to the HEW Training Institute on Evaluating Clinical Performance, Kansas State University, Manhattan, Kansas, April 1980.

McKeachie, W.J. "Psychology in America's Bicentennial Year," *American Psychologist,* vol. 31, 1976, pp. 819-833.

Ory, J.C. and L.A. Braskamp. "Faculty Perceptions of the Quality and Usefulness of Three Types of Evaluative Information," *Research in Higher Education,* in press.

Steers, R.M. *Organizational Effectiveness: A Behavioral View.* Santa Monica, Calif.: Goodyear Publishing, 1977.

NSPER: 80

Co-Directors' Presentations

ROBERT B. INGLE
University of Wisconsin
at Milwaukee

WILLIAM J. GEPHART
Phi Delta Kappa

Techniques for Systematic Observation

Robert B. Ingle

Observation, as a technique for evaluating teachers, has a long and checkered history. In most instances it went something like this.

The day was bright and sunny. You were in your classroom teaching for all you were worth to all of those eager seekers-after-knowledge when suddenly the door was thrown open and in strode a PERSON OF AUTHORITY—the superintendent, the principal, a supervisor, or what have you. (At about that time, clouds would usually darken the sun and the day would take on a threatening character.) You looked up, startled, and were told: "DON'T LET ME DISTURB YOU; GO RIGHT ON WITH WHAT YOU WERE DOING." (It always seemed that THEY spoke in capitals, but perhaps that was because we were young.) In any event, the person of authority took a seat at the back of the room, took out some writing material, and proceeded to watch.

If you were lucky, the person of authority watched you with a benign look on its face. (I have trouble attributing anything but a neutral gender to them.) If you were not lucky, it watched you with a malevolent look on its face. The creature at the rear of the room might occasionally make a note or, worse yet, write furiously. Whatever happened, the effect on you was likely to be somewhat unnerving at best and catastrophic at worst. Your ordeal might last for the entire class period or, with no warning, the creature might heave itself out of its seat and shuffle out of the room. Whether the creature would descend upon you again and, if so, when, was anyone's guess. Although possibly not calculated to, the occurrence and accompanying uncertainty for the future left you in a highly nervous state, a state not likely to have a positive effect on your teaching.

What I have just described, in overly dramatic terms perhaps, is known as unstructured observation. The individual being observed did not know what the observer was looking for, and the observer, except in the most general sense, did not know what he or she was looking for.

The usual rationale offered for this type of observation was that the observer had more years in education, thus more ex-

perience, and "knew what to look for to decide whether the teaching was good or bad." The fact of the matter is that the observer may have had some internal, hidden criteria, but no one necessarily knew what they were. In most instances the observer would probably have had a difficult time verbalizing his or her specific criteria for judging good teaching. And, if the observer could describe his or her criteria, you might discover that you disagreed with them.

The value of such observations is minimal. Two observers may agree on the quality of a teaching performance, but for quite different reasons. And, two observers might come up with different estimates of the quality of a teaching performance. For obvious reasons, I reject the use of unstructured observations for the evaluation of teaching—no matter how experienced and intuitive the person of authority may be.

Observation can be used, however, in a systematic and structured fashion. In order to use observation in this way, it is necessary that the observer have a guide, a structure to follow while doing the observation and that the observer confine himself or herself to the guide.

Checklists and rating scales are the two most commonly used guides for observing teaching. The former consists of a listing of steps, activities, or behaviors the presence or absence of which are to be observed. The task of the observer is to indicate on the checklist whether these steps, activities, or behaviors did occur. A checklist might look something like the following, although I don't present it as a model:

Activity	Yes	No
Teacher explained purpose of lesson	_____	_____
Teacher presented lesson in an organized fashion	_____	_____
Teacher reviewed lesson at the end of the class period	_____	_____

It should be emphasized that a checklist enables the observer to note only whether a given behavior occurred. It does not ask the observer to rate the quality, degree to which, or frequency of occurrence of a particular behavior. When such information is desired, a checklist is definitely inappropriate. Since the checklist only allows a recording of whether a particular behavior took place, it provides only part of the information needed in the evaluation of teaching. A checklist would be appropriate when the presence or absence of certain activities or behaviors have been

identified as crucial in evaluating teaching. Even then, it is impossible or at least inappropriate to do more than note their presence or absence.

Somewhere between a checklist and a rating scale (which will be discussed shortly) is a guide that allows the observer to record the number of times a given event occurs. An example of such a guide is the Flanders Interaction Analysis Observation System, an example of which is shown on page 110.

This type of observational guide provides a description of what occurs and the number of times it occurs. Considered alone, such a guide does not provide any indication of the quality of the behaviors observed. In the use of such a guide it is tempting—and many have succumbed to the temptation—to make judgments about the worth of various activities, or classes of activities, with little or no evidence that any of the activities contribute to good teaching or that one activity is better than another. Such a guide facilitates description, but not value judgments—at least not without additional data.

Another form of guide for observing teaching is the rating scale. It not only allows the observer to indicate whether the behavior occurred, but also to make a qualitative judgment about the behavior or activity.

In general there are two types of rating scales. One type contains a numerical key or scale that is used for rating all attributes to be observed. For example:

5 — excellent
4 — above average
3 — average
2 — below average
1 — poor

For each attribute or teaching behavior, the observer either checks the appropriate box or circles the appropriate number. This type of scale is useful if the key is appropriate for all of the attributes being rated—usually not the case. Some try to create rating scales in which the same scale is used to rate the activities listed on a checklist. These general descriptors used for the key often do not fit some of the activities listed. In such instances the rating scale needs descriptors that are specifically developed for each activity. This is particularly evident if, for example, the behavior to be rated was "Teacher called each student by name." An excellent-to-poor scale does not fit here. In such a diversified activity as teaching, it is difficult to find a single key that is appropriate for all attributes that one might wish to rate.

The other type of rating scale is a graphic rating scale. On this type of scale the observer can check anywhere along a continuum; he or she is not confined to any particular point. Although this gives the rater greater freedom, it can prove a problem if one is trying to get a score for the various attributes. This type of scale is preferred when there is no common key appropriate for all of the attributes being rated. Using the same activities as were previously used, a graphic scale would look like the following:

Teacher explained purpose of the lesson

1	2	3	4	5
Did not explain		Explained adequately but not fully and completely		Explained fully and completely

Some of the *advantages* to the use of systematic observational techniques are:

1. They are directed toward specific aspects of classroom behavior.
2. They provide a common frame of reference for comparing all teachers on the same set of dimensions.
3. They are a convenient method for recording appraisals by the observer.
4. They facilitate communication between observer and observee.

Some of the *limitations* in using systematic observational techniques are:

1. Unless considerable care is taken in construction, the meaning of various scale points can be ambiguous—the term "excellent" is an example.
2. They are subject to various observer biases:
   A. halo effect
   B. leniency or generosity error
   C. severity error
   D. central tendency errors.
3. There may not be much evidence that the attributes observed relate to student achievement.

Interaction Analysis Observation System*

With this type of format, the observer counts separate behavioral incidents. For example, if the observer is interested in a teacher's questioning behavior, the observer counts the number of teacher questions made with the intent that a student answer.

TEACHER TALK	INDIRECT INFLUENCE	1.** ACCEPTS FEELING: Accepts and clarifies the feeling tone of the students in a non-threatening manner. Feelings may be positive or negative. Predicting or recalling feelings are included. 2.** PRAISES OR ENCOURAGES: Praises or encourages student action or behavior. Jokes that release tension, not at the expense of another individual, nodding head, or saying, "um hum?" or "go on" are included. 3.** ACCEPTS OR USES IDEAS OF STUDENT: Clarifying, building, or developing ideas suggested by a student. As a teacher brings more of his own ideas into play, shift to category five. 4.** ASKS QUESTIONS: Asking a question about content or procedure with the intent that a student answer.
	DIRECT INFLUENCE	5.** LECTURING: Giving facts or opinions about content or procedure; expressing his own ideas, asking rhetorical questions. 6.** GIVING DIRECTIONS: Directions, commands, or orders to which a student is expected to comply. 7.** CRITICIZING OR JUSTIFYING AUTHORITY: Statements intended to change student behavior from nonacceptable to acceptable pattern; bawling someone out; stating why the teacher is doing what he is doing; extreme self-reliance.

STUDENT TALK	8.** STUDENT TALK--RESPONSE: A student makes a predictable response to teacher. Teacher initiates the contact or solicits student statement and sets limit to what the student says. 9.** STUDENT TALK—INITIATION: Talk by students which they initiate. Unpredictable statements in response to teacher. Shift from 8 to 9 as student introduces own ideas.
	10.** SILENCE OR CONFUSION: Pauses, short periods of silence and periods of confusion in which communication cannot be understood by the observer.

-111-

*Flanders, N.A. *Teacher Influence, Pupil Attitudes, and Achievement.* U.S. Office of Education, Cooperative Research Monograph No. 12. Washington: U.S. Government Printing Office, 1965.

**There is NO scale implied by these numbers. Each number is classificatory; it designates a particular kind of communication event. To write these numbers down during observation is to enumerate, not to judge a position on a scale.

Developing Rating Scales and Checklists

Robert B. Ingle

At this point the focus turns to items that should be considered when choosing or developing a rating scale or checklist. They are generally relevant to all of the forms of observation previously discussed; some, however, are particularly relevant to rating scales.

1. *The items chosen should be demonstrably related to effective teaching.* From what we have heard from the research reports on teacher effectiveness, this may incline everyone to give up the task and go home. There are, however, some traits that appear to be related to effectiveness—teacher enthusiasm, for example. Other sections of this report discuss some of those traits.

2. *For items that fall essentially into the realm of values, general agreement should be reached between those being observed and those doing the observing about the inclusion of such items.* General agreement between the parties should be easy to reach on those items that are demonstrably related to effective teaching. That is, agreement is not as difficult to reach when the items under consideration are consistently associated with student growth in the research on teacher effectiveness.

When we enter into the realm of values, the situation is significantly different. It is easy to argue that my values are as good as your values—maybe better. Thus it follows that I (as the person being observed) should not be penalized by a poor rating for not having a pretty bulletin board display just because you (as the observer) *value* pretty bulletin board displays. I do not value them at all. The research literature is ambivalent on this and similar items. Some studies conclude that they are associated with student achievement, while others conclude that they are not. The justification for including them is that we, as a staff, value such an activity/accomplishment as an aspect of teaching.

There are at least two types of value-laden items that may get included on rating scales and used in evaluation of teaching. One type includes items that appear to be logically related to effective teaching, although immediate evidence that an item does, in fact, relate to effective teaching is inconclusive. For example, evidence

gathered in studies on animal learning is often extrapolated to human learning. It is reasoned that if too much anxiety can interfere with animal learning, then it is likely to also interfere with human learning. Many items classified as relating to classroom climate would fall into this category.

The other type of items are those that, even logically it is hard, maybe impossible, to argue should be related to teaching effectiveness. An example of one such item would be teachers' dress. Another, I would argue, would be pretty bulletin board displays. I do not mean to suggest that individuals will not try to include such items; they will, because they value them. Their arguments for including them are tortuous, however, and do not stand up well in the clear light of reason.

The point is that the inclusion of value-laden items, ones that are not demonstrably related to effective teaching, even such an item as the prettiness of a bulletin board display, is allowable given staff consensus on the inclusion. Obtaining such agreement is crucial.

3. *Include only behaviors or activities that are observable in the teaching situation.* This point should be obvious, but it is better to make it explicit rather than to hope it is implicit. Those developing observation instruments should ask themselves regarding each item whether the behavior or activity that is to be rated is observable and is likely to be elicited in a normal classroom situation.

4. *Use low-inference rather than high-inference items.* An example of a low-inference item is "teacher smiled." A high-inference item is "teacher was enthusiastic." You would get little disagreement as to whether a teacher smiled. You might get much disagreement as to whether a teacher was enthusiastic. In terms of the previous point about observability, you are much more likely to observe a behavior called for in low-inference items and get agreement on their presence or absence than you are in behavior called for in high-inference items. There is nothing wrong with using a high-inference item, such as teacher enthusiasm, as a starting point. They are particularly useful for developing low-inference items. Teacher enthusiasm, for example, could be broken down into a series of simpler behaviors that could be directly observed. Those involved in the evaluation would have to agree that the various simple behaviors (low-inference items) are related to teacher enthusiasm, of course. To the extent that such agreement is reached, you will have a much better measure of teacher enthusiasm. A dissertation done at Syracuse University in 1976 by Mary Collins is illustrative. She breaks teacher enthusiasm into eight low-inference behaviors.

5. *Clearly define the traits to be rated and the scale points to be used.* To the extent that low-inference items are used, the traits are more likely to be clearly defined. The scale points are quite a different matter. For example, suppose you were asked to rate the occurrence of a given behavior and the scale points were:

5 — very often
4 — often
3 — frequently
2 — infrequently
1 — rarely

How much more than rarely is infrequently? How much more than frequently is often? For that matter, how much is frequently, often, or any of the others? Without further definition, the rater is left up to his or her own devices to decide. The likelihood of agreement among raters and between raters and those being rated in such a situation is small. It would be much better in this situation to quantify the scale points. For example, "very often" might be 90% + of the time; "often" might be 75-90%; and so on. In this way you are more likely to get those using the scale to make the same judgments. For each scale point, the developers should ask themselves what they mean by that point and define it as clearly as possible.

6. *The scale points must be appropriate for the item being rated.* When the item calls for essentially a frequency measure, as in the previous point, the problem is not too great. If, however, the scale points essentially define the item, there may be a problem. Consider the following item:

Enjoyable Environment: Does the teacher make provisions for an enjoyable lesson?

1	2	3	4	5
Students are always serious		Students occasionally smile and laugh		Students frequently smile or laugh

Note that an enjoyable lesson is defined by the scale point statements. One must assume that whoever developed this item considers an enjoyable lesson one that makes students laugh or

smile frequently and that a lesson is less enjoyable if the students are always serious. I am not at all sure that everyone would agree with that definition of enjoyable lesson. The item is clearly a value item, although a logical argument based on motivation theory and learning environment studies might be made. The definitions used for the several scale points are another matter. Whether they define the continuum of the enjoyability of a lesson is not at all certain.

7. *Do not ask for more discriminations than raters are able to make.* The number of discriminations is determined by the number of scale points. On the previous item, the rater is asked to make at least five discriminations, although two are not defined. If you cannot meaningfully define each point on the continuum, you probably ought not to use that many points. If you cannot indicate what the nature of the discriminations are, it is unlikely that a rater will be able to make the discriminations. If, on the previous item, a teacher got an average rating of 3.8, how would you categorize it? About all you could say is that the teacher is nearly halfway between occasionally and frequently, but how that translates into smiles or laughs is unclear.

In addition to not asking raters to make more or finer discriminations than they are able to, you should not ask raters to make any more discriminations than will be used or that are meaningful. For example, if it is necessary to make a global rating of a teacher's ability you ought to ask yourself if you really need more than two categories—satisfactory and unsatisfactory. If the answer is yes, you should ask yourself why. If you can think of a good answer, then you should add more categories. I can think of very few instances when you would need more than two categories. Of course, if you were using the evaluations for the improvement of instruction you would not be likely to have such an overall, global rating.

8. *Avoid extremes at the ends of the rating scale.* The attributes upon which you rate teachers are generally not characterized by "always" or "never," or those conditions occur so infrequently that it is not worth using the terms. If "rarely" is substituted for "never" and "usually" substituted for "always" (all terms suitably defined, of course), you will find the scale much more useful. In the item on enjoyable environment, if "usually" is substituted for "always" at scale point one, you would improve the scale, since the number of classes in which students are always serious is few. The use of extremes at the ends of your scale has the effect of reducing your effective scale by two points, because the extremes are rarely checked.

9. *With a rating scale, use a total score only if the items on the scale are exemplars of a higher order attribute.* For example, if teacher enthusiasm is broken down into a number of specific behaviors, it may make sense to add the points on each item for a total score that represents teacher enthusiasm. On the other hand, if each item on the scale represents a distinct attribute, it makes no sense to add the scores on each item to obtain a total score, unless the various items are compensatory, i.e., a low score on one item may be compensated for by a high score on another item. If each item represents a separate and distinct attribute, a total score is meaningless.

Observation is a data generating process that has been used in systematic inquiry in many areas and for a long time. In recent years its use in evaluation of teaching has been called for in greater intensity than in prior decades. The use of observation for evaluating teaching demands that people who create observation instruments have a clear understanding of the fundamentals of item selection and scaling. The guidelines above are presented toward this end.

Components of a Humane System for Evaluating Teaching

William J. Gephart

Many forms of measurement have been employed to assess the quality of the teaching that takes place in a given school or classroom. We can lump these various forms of measurement into two categories. The first includes those measures that focus on the teaching process—what does the teacher do? The second category includes those focused on product—what students can do supposedly as a result of that teaching. The NSPER:80 sessions, following a course started a few years earlier, focused on the improvement of teaching through improved process measurement, primarily the observation of instruction.

There are four different types of observation of teaching that can be seen in educational practice. Some observations are unstructured; someone comes into a classroom, sits down, and watches for some period of time unguided by any specified set of actions they wish to observe.

If there are specific activities for which they are looking, we would classify that observation as a structured observation. For example, an observer might have a checklist of teaching activities that are to be watched for during the course of the observation. The clear statement of the instructional goal being focused on, checking to be sure students are doing assigned seatwork correctly, demonstrating enthusiasm for learning the subject matter, etc., are illustrative of the kinds of items that might be contained on such an observation checklist. Two general forms of structured observations have been used, the qualitative and quantitative approaches. The latter asks how often a given teaching activity occurred. The former asks about the quality of the performance.

The fourth form of observation is one in which, either by words or pictures, we give a description of a classroom situation that presents an instructional problem. We then ask individual teachers, singularly or in groups, to observe or to read the description of the situation and to report on what they would do if they found themselves in that situation. Their response to the situation is then scored using a list of preset criteria. That score is the indicator of the quality of the instruction for that person.

In the past several years a great deal of attention has been

focused on observing instruction. In some states systematic observation of teaching has been mandated by legislative or state department of instruction action. Because of statements by many NSPER participants regarding these mandates and their concern for improving education, the planning for NSPER:80 was focused on the improvement of structured observation. The program for the three 1980 sessions was designed to identify the guidelines and principles that educators believe should be followed in the construction of observation instruments. As the planning progressed, it became clear that the focus was *not* on developing THE observation instrument that would resolve all educators' problems. Rather the focus was on what might be called "use-tailored" observation instruments. We would not expect to see the same instructional activities carried out in a kindergarten and a high school trigonometry class. Neither would we expect teachers of music and of foreign languages to engage in exactly the same activities. Given those differences and the desire to use observation as one technique in our efforts to evaluate and improve instruction, we need to tailor our observation instruments so they can deliver the kind of information needed for the evaluation of teaching. The focus of NSPER:80, then, was on the process for *developing* observation instruments (note the plural), not on the creation of *an* instrument that would be used everywhere.

Contexts for the Evaluation of Teaching

Economic factors and declining enrollments have enhanced the importance of evaluating teaching. Over the past five to 10 years a shortage of teachers has changed to a surplus. Teaching jobs, once easy to find, are now very scarce. As a result teachers who get a job tend to stay in that place for longer periods of time. Teacher mobility has reduced markedly. Gone are the days in which a person would teach in one location for a while and then easily find employment in a different section of the country. Given less mobility, one source for new information and new teaching approaches—the hiring of new people—has been all but eliminated. In those circumstances we have to find new ways of promoting professional growth in a stable faculty. Evaluation of the quality of teaching is being undertaken as one component in a school system's response to these problems.

The demand that teaching be systematically evaluated is not without major problems. Evaluation produces anxiety.

Educators, both those doing evaluations and those being evaluated, need to accept that as a given. And, we need to recognize that some anxiety is needed for progress and that too much anxiety is destructive to the accomplishment of our tasks. The trick is to control the anxiety, to keep it within those bounds—a trick that is easy to call for and difficult to accomplish. If you are an evaluator and I an evaluatee, I will picture you somewhere on a continuum from helpful and supportive at one end and threatening on the other.

A second problem in meeting the growing demand for evaluating teaching is uncertainty about the process itself. How can personnel performance be evaluated? Who should be involved? What techniques should (could) be used? How should the results of evaluations of teaching be used? What are the rights and responsibilities of the individual being evaluated? Of the institution? Much has been written about the *how* of performance evaluation, but many people are still uncomfortable about it, both those who are being evaluated and those who must do it.

Finally, there is the time problem. Educators are busy people. The demand for evaluation is a demand for time expenditure. There is no way that a one-hour observation of teaching can be done in less than 60 minutes. Where will the time come from? What current activities can be eliminated to meet the mandate for evaluation? Will the evaluation process reduce the time requirement for some of those other activities?

The evaluation of teaching is done for a number of different purposes, two of which are so central to the work of NSPER:80 that they need to be described here. Those two purposes can be seen in the following personalized statements. Sometimes we evaluate my teaching to give me information I will use in my decisions about how I will teach the subject matter I have been assigned to teach. At other times we evaluate my teaching to give administrators and supervisors information they can use in their decisions about how they will use me. The first of those approaches is evaluation for the purpose of professional growth. Evaluation for professional growth attempts to provide information that a teacher needs to make decisions about his or her work style and professional competencies.

The second approach provides information that is necessary in the operation of the organization. The people responsible for the operation of that organization or that institution must continually make decisions about staffing. They are faced with decisions about employment, promotion, retention, and in some cases even pay.

Throughout their deliberations the participants in NSPER:80 were constantly brought back to the point that the purpose for evaluating teaching was to facilitate decision making either in the direction of professional growth or in the direction of managerial decisions. The question of purpose is primary. The purpose being served should always be clear to those persons who create and/or use observation instruments. It should also be clear that observation and evaluation are not synonyms. They are not identities. When we observe we generate descriptive data about the teaching that is going on. That data is used along with other kinds of data in evaluating teaching—i.e., the development of judgments about the quality of the work of a teacher.

The participants of NSPER:80 referred from time to time to the efforts in many locations to develop an instrument to guide the observation of instruction. In some cases these discussions focused on a state effort, in others to a school systemwide evaluation. They are too general and abstract and, as such, are not helpful in specific cases. As these discussions progressed it became clear to the participants that the information generated by such global instruments are "high inference" data: "the teacher created an enjoyable learning environment," for example. The discussions recognized that, to help in the improvement of student learning, the observations needed to focus on specific things the teacher does. These discussions combined with the earlier concerns about teaching at different levels and different subjects led to general agreement that an instrument could not be developed that could be used across all levels and subject matter instructional tasks. To develop an encompassing instrument it is necessary to state the things to be observed in such general form that they would lack specificity and prescriptiveness. They do not create action-specific information for reporting to teachers about what *might* be done to improve instruction. This discussion leads again to the idea of a "use-tailored" observation instrument. Consensus seemed to develop around the idea that we need multiple instruments, instruments developed for creating as specific information as we can possibly get. An evaluation that will help an individual improve as a teacher needs information that speaks particularly to instruction at the kindergarten level, or the primary level, etc. And, we need information generated by instruments that focus directly on the teaching of mathematics or language, etc. As stated above the NSPER:80 groups began to refer to these as "use-tailored" observation instruments.

It is within this problem-laden context that the NSPER:80 discussions took place. The NSPER:80 participants recognized

that at this point there is no one right way to evaluate teaching. In light of these circumstances the discussions of NSPER:80 concentrated on guidelines and principles that should be followed. The remainder of this wrap-up focuses on those guidelines and principles.

Principles and Guidelines for Developing Use-Tailored Observation Instruments

The push to improve the evaluation of teaching through systematic observation has caused administrators to search for "a good instrument." It is not uncommon to hear a group of school administrators talking about their satisfactions and dissatisfactions with the observation instruments they are using. It is also not uncommon for the people who are dissatisfied with their observation instruments to ask those who are satisfied if they might have a copy. The NSPER:80 participants believed that administrators should search for exemplary instruments. However, the NSPER:80 participants do not believe that school system A should merely adopt the instrument believed successful in school system B. An observation instrument is constructed on a value base that defines what good teaching is or is not. Research on the components of effective teaching (an activity that has been underway for eight or nine decades) is inconclusive.

In our research there are four or five items that consistently and convincingly correlate with student achievement and growth. Those items are: knowledge of the subject matter being taught, articulateness of the teacher, enthusiasm displayed by the teacher for learning about the subject, time-on-task structured by the teacher (some use the phrase engaged time), and the establishment of an orderly, business-like atmosphere. Few people will accept those items as a complete description of "good teaching." Some will insist that other items are just as important, or more so, than those found in research. For example, some people feel (believe) that a good teacher must have positive regard for the learner. Our knowledge is inconclusive on this point. Others believe that the way a person dresses is a component of teaching. Again, our knowledge is inconclusive. These items and others are the expression of the values of people. We know that values differ from place to place. One school's definition of teaching—and thus the structure for its observation instruments—will reflect the values in that school and community.

The NSPER:80 participants strongly believe that observation instruments should be based on the value system that exists in the school or schools in which the observation instrument is to be used.

To base an observation on the school/community's values mandates an activity. Each school system needs to create and maintain a policy statement that has two parts. The first part is a carefully constructed consensus statement that defines what teaching is in that school system and what it is not. Without such a consensus statement there is no basis for appraising teaching. If we do not agree on the components of something to be done, we cannot appraise the quality of what is done. The policy statement should be an operational description of teaching, what it is, how it works, and what its anticipated impacts will be. It should not be a grand and abstract statement that occupies file drawer space.

The second part of the policy statement should be an operational statement descriptive of what evaluation is in the specific school system. Again the NSPER:80 discussions called for creation and maintenance of an operational policy. That call is a recognition that teaching—and thus the evaluation of it—is not a static entity. As society's goals change, teaching and schooling should also change. What an evaluator of teaching would have looked for in a classroom observation in 1932 is different from what would be looked for in 1982.

There is a second reason that our policy statement defining teaching and its evaluation should be conceived as dynamic entities: It is highly unlikely that we can create the perfect evaluation plan the first time. We should recognize that evaluation plans and instruments are likely to have flaws; we should expect that changes in instruments and procedures will be made. The personnel performance evaluation plan has a purpose, or purposes. We should evaluate the degree to which the plan is enabling the achievement of that purpose and make changes as deemed necessary.

The second major principle underlying the creation of use-tailored observation instruments is involvement. It is important for people at all levels—those being evaluated, those for whom the evaluation is being done, and those doing the evaluation—to be involved and/or contribute to the creation of the instruments. Of course, in large school systems everyone cannot be involved. However, care should be taken to involve representatives of all types of people who will be involved with the evaluation. Involvement creates ownership. If the observation instruments to be used in the evaluation are created by one group alone, it can

be extremely difficult to get other groups to accept the information produced by those instruments. Any flaw in the instruments created by "that group" will be hard to overcome. Rather, it will be one more illustration that those "experts" really don't know what is going on in the schools. Ownership is an important ingredient. If our system has a flaw, it is much more likely that we will retain our commitment to the system AND search for ways to correct the flaw.

The recommended involvement of people from all levels has another basis. A depiction of teaching will vary as vantage points change. Teaching as perceived by a teacher will be different from a depiction by students, or supervisors, or administrators, etc. In developing the operational policy statement, input from these various vantage points should be sought. The same holds for creation of an administrative performance plan. Teachers have a view of administrator performance which should not be overlooked in planning and carrying out evaluative tasks. Involvement of all the people, the evaluators and evaluatees, is an important principle in the creation of observation instruments tailored to create data that will be useful in both professional growth and managerial decision making.

An examination of the forms used in school systems for the evaluation of teaching will produce a variety of items on which the evaluation is to focus. In many, many instances a large number of those items are not directly observable in the classroom. The NSPER:80 sessions restricted attention to those things that can be observed in a class. This is not a denigration of the value of other kinds of data in the evaluation of teaching. Rather it is an effort to sharpen the focus, to concentrate attention on observation of teaching in the classroom. People responsible for the evaluation of teaching should recognize this limitation and should create other instruments for generating data about characteristics or activities that are not likely to be seen in the classroom actions of teachers. Previous NSPER sessions (Planning for the Evaluation of Teaching—NSPER:79, Phi Delta Kappa, Bloomington, Ind., 1979) identified four different classes of variables (academic achievement and physical, moral, and emotional growth); teaching style variables; employee variables (honesty, dedication, ability to work with others, etc.); and, context variables (the constraints on teaching that come from state actions, local board regulations, operational procedures in a given school, and the quality of students assigned to the teacher). In evaluating teaching all of these categories of variables should be considered. Obviously, many of these kinds

of data can not be generated through classroom observation. Consequently, an evaluation plan is needed that delineates the kinds of data needed for the different kinds of decisions, how they will be collected and used. Classroom observation is but one component of that evaluation plan.

The participants expressed concern about high- and low-inference items that show up on instruments for the observation of teaching. Low inference items on an observation guide or checklist are directly observable. For example, teacher stated the objective for the day's lesson. That is something that can be directly observed. It does not require stringing together a number of observable items in some logical inference pattern. An illustration of a high-inference item is teacher enthusiasm. A number of different things would be observed to determine a teacher's enthusiasm. To as great a degree as possible the focus of an observation instrument should be on low-inference items—those things that can be seen directly rather than those which must be inferred.

Creating Use-Tailored Instruments

Research on the effectiveness of teaching has been underway for approximately eight or nine decades. In that time a lot of different factors have been explored in an effort to determine what the components of effective teaching are. The major resource people for NSPER:80 indicated that in the past several years there have been a number of attempts to review that research to determine what we know about the effectiveness of teaching—i.e., to identify the fundamental characteristics of effective teaching

Those summaries taken together identify four different factors, or aspects of teaching, that consistently correlate with student performance. They are: 1) the knowledge of the subject matter, 2) articulateness of the teacher, 3) the teacher's enthusiasm for learning about the subject he or she teaches, 4) time on task (engaged time).

The research is conclusive that if two teachers whose knowledge about the subject matter differs are assigned to teach that subject matter to comparable students, the students with the teacher who possesses the greatest amount of subject matter knowledge will typically outscore the students of the other teacher.

The same holds for the articulateness of the teacher. If two teachers differ in their ability to communicate about the instruc-

tional procedures or about the subject matter, the teacher who is best able to articulate will have students who perform better on tests over that subject matter. (This factor of articulateness has been explored by Donald Cruickshank and John Kennedy at Ohio State University. Their work and that of several others has been pulled together in the June 1981 issue of Phi Delta Kappa's newsletter, *Practical Applications of Research.*)

The teacher's enthusiasm for learning the subject has also been explored at some length (see the June 1981 issue of *PAR*). The focus here is not on general enthusiasm, but rather on the degree to which the teacher shows excitement about learning the subject matter she or he teaches. Is the teacher an ongoing learner related to the subject matter? Or is this a job, a job that needs to be performed and is carried out in a rather lackluster way. There are a number of different ways that enthusiasm can be displayed. Again the research seems conclusive; the more the teacher displays enthusiasm for learning the subject matter, the higher that teacher's students will perform on tests on that subject matter.

The fourth item, time-on-task or engaged time, is a relatively new focus in the research literature. Its incidence during the past decade, however, and the nature of the findings, show with some conclusiveness that the teacher's ability to structure learning time, time that includes specific learning activities to be engaged in by the students, is an important factor in effective teaching. It is a move away from concern about allocated time and toward concern for expended time. Debates raged for some time in our field regarding how much time should be allotted to the teaching of reading. The time-on-task research focuses on how the time is spent and the degree to which the teacher can structure students' learning time. Again, time-on-task correlated highly with student performance.

The four different factors, despite the fact that they consistently correlate with student performance, are not likely to be accepted by anyone as a complete definition of teaching. There are other items that individuals and groups among us would like to see included as we attempt to define what effective teaching is. These other items can best be referred to as *hypothesized* corrolaries of effective teaching. They are things that individuals among us *feel* are important aspects of teaching. At the same time they are not supported nor refuted by our accumulated research literature. For example, some people strongly believe that a teacher must show positive regard for the learner; some would say unconditional positive regard. The research literature is not conclusive on that point. Some studies support the conclu-

sion that the teacher who displays unconditional positive regard will have consistently higher performing students; other studies fail to support such a conclusion. Another illustration is the creation of an enjoyable environment. Again, the research literature neither supports nor refutes that item. Regardless of the support or lack of support in research, we *need* to consider these hypothesized correlates as we create instruments for observing teaching. In effect, these hypothesized correlaries are values—items that are held dear on the part of numbers of people in a given setting. As such they are a part of "what is important" in the given setting and *will* play a role in assessing the quality of teaching there.

The participants of NSPER:80 firmly believe that we should take direct steps to include those items in the observation of teaching. Valued hypothesized components of teaching will play a role in the perception of the quality of teaching. Often we are only semiconscious of that role. It is, according to the NSPER:80 participants, much easier to deal with something that is out in the open, on top of the table, rather than in a hidden agenda format.

It is extremely important that the staff in a school come to consensus about the components of effective teaching—what does "teaching" consist of in *our* setting—if they are to have success in creating useful observation instruments. Achieving that consensus is not easy. In requires time, time that some of us would rather spend teaching. We need to clearly understand the research AND the value-based items that can only come through widespread discussion that illuminates variation of perception. The NSPER:80 participants believe that the time invested in talking about what we feel, what we value as the components of teaching, is well worth the expenditure of money it will entail.

The first step in the creation of an observation instrument is the development of the list of the components of teaching that come from the staff consensus referred to above. To get that list, either all the staff or representatives of the various elements of the staff should be scheduled for a work session in which they first talk through the research-based components of effective teaching. The purpose of this discussion is the achievement of clarity about what is and isn't included in the four research-based correlates of effective teaching. Next, the staff (or the selected representatives) should be asked to list what is to them *the* comprehensive set of items that comprise effective teaching. This list will contain both research-based and value-based items.

At this point a concern must be broached about the level of abstraction in the discussion of the components of teaching.

Conceptually, it would be possible to list almost every specific act engaged in by the teacher (e.g., takes attendance, collects and turns in milk money, records grades, etc.). It is also possible to talk in great generality (e.g., presents a lesson, diagnoses needs, etc.). It is strongly suggested that in the initial stages of creating an observation instrument that the attention focus on more general terms, more general classes of activity. The goal here is a list of eight to 12 items, not a list of several hundred items. This may sound as if it were in conflict with the idea about being specific for the level or subject matter being taught. It is not. Rather it is a recognition that our understanding of teaching is not a precise science. We do not know that the addition of a pound of this or an ounce of that raised to so many degrees will create the desired effect. Teaching is much more of an art. It seems to involve a number of things, but exactly how those fit together we are not prepared to say at this point in developing understanding of our field. Thus to focus on each specific teaching action would be misplaced precision.

There are other reasons for opting for listing of a more general category. The evaluation of teaching takes place in a limited time frame and in competition for a limited quantity of resources. Therefore, we cannot expect to focus on everything. Rather, we need to pick those things that are important to us as the major things to attend to in our efforts to improve instruction.

Given the development of such a list the next step is the determination of the priority of those items. Most observation instruments that we look at, in fact most evaluation forms used in schools, include a number of different items with no indication of differences of importance. The assumption seems to be that every item listed is of equal importance. That seems to be an untenable assumption, one that needs to be tested out. This can be done relatively fast by involving the people in a paired comparison analysis to determine the priorities, relative weights of the different items in their listing of components of effective teaching. The paired comparison technique asks people to compare every item on the list with every other item on the list. Analysis of the preferences expressed by the respondents shows that perceived ranking of the items being considered and the degree of importance within that ranking.

After the relative value of the different items has been determined the people creating the observational focus have an indication of where the staff feels the observational focus should be if it is to be of greatest utility in improving instruction. That

priority listing can be used to determine where a staff wants to devote its attention. That determination should be made by a representative group. But, since it is primarily improvement of the institution's effort that is our focus, the administrative and policy makers ought to have the greater decision-making responsibility.

When the five to 10 top priority items are selected, we need to engage the staff again. This time we should ask them to determine for each of these items the low-inference observables that we could list on our instrument. What are those things that we would expect to be able to see if "good teaching" is happening in the classroom? What are the things that would be negative indicators, teacher actions that are unacceptable? Those are the items that we should list on the observation form. This discussion may identify a number of observable activities for each of the priority items. Classroom management skill might be inferred by seeing: 1) whether instructional materials are available, 2) whether the teacher gives directions clearly and explicitly, and 3) whether the teacher monitors assigned activities, etc.

The structure of the form is another consideration. If the data we believe we need is qualitative in nature, we need to be sure that we state the dimensions of that quality and label the levels that we are expecting to see. If the data we need is frequency data, then we need to provide space for the observer to record that frequency information.

The form should be structured physically so that items that are of a similar nature are placed near each other to expedite the observation itself. The instrument should also group those items that are believed to be compensatory in the system's definition of good teaching. If being outstanding in teaching activity A makes up for (or compensates for) being low on teaching activity B, a compensatory relationship exists and the form should expedite interpretation of that relationship. Other activities are independent criteria. For example, we want every staff member to be honest. We seldom accept more or better performance in some other area in lieu of honesty.

Finally, before the newly developed observation forms are used, time should be taken to acquaint the entire staff with them. The goal here should be, "No surprises." Everyone should know what the instruments are, how they are to be used, what they produce, and what will be done with those products. Evaluation produces anxiety. We need to avoid surprises to help control the anxiety that accompanies the evaluation of personnel performance in our schools.

Summary

The three NSPER sessions in 1980 focused on principles and guidelines for the development of instruments for observing teaching. The individuals who participated believe that such instruments should be "use-tailored," that is, they should be created to be used at the level and in the subject matter on which the observation is to be focused. They believe strongly in the involvement of people representing all the different levels of education in the development of the observation instruments for their teacher evaluation efforts. Observation is one way to generate information about the quality of teaching. It is a process-oriented measurement form. It produces data that can be useful to a variety of educational decision makers.